Topics in quasi-Newton and space-time spectral methods

by

Sarah Nataj

A thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

Doctor of Philosophy

Department of Mathematics

University of Manitoba

Winnipeg

Abstract

The first part of this thesis focuses on quasi-Newton methods. Broyden's method is a quasi-Newton method which is used to solve a system of nonlinear equations. Almost all convergence theory in the literature assumes existence of a root and bounds on the nonlinear function and its derivative in some neighbourhood of the root. All these conditions cannot be checked in practice. The motivation of this work is to derive a convergence theory where all assumptions can be verified, and the existence of a root and its superlinear rate of convergence are consequences of the theory. The theory is simple in the sense that it contains as few constants as possible. The method of Broyden-Fletcher-Goldfarb-Shanno (BFGS) is also a quasi-Newton method for unconstrained minimization. Also, all known convergence theory assume existence of a solution and bounds of the function in a neighbourhood of the minimizer. We generalize a convergence theory where all assumptions are verifiable and existence of a minimizer and the superlinear convergence of the iteration are conclusions.

In a continuation of this part, we consider Perry nonlinear conjugate gradient (NCG) method and scaled memoryless BFGS method. These methods represent important schemes for solving large-scale unconstrained optimization problems. Only the basic versions of these methods without line search are considered. We show local superlinear convergence assuming hypotheses which can be verified in practice.

In the second part of this thesis, space-time spectral methods are considered. Spectral methods solve ordinary differential equations (ODEs) and partial differential equations (PDEs) numerically with errors bounded by an exponentially decaying function of the number of modes when the solution is analytic. For time dependent problems, almost all focus has been on low-order finite difference schemes for the time derivative and spectral schemes for spatial derivatives. Spectral methods which converge spectrally in both space

and time have appeared recently. In this thesis it is shown that a Chebyshev spectral collocation method of Tang and Xu [71] for the heat equation converges exponentially when the solution is analytic. We also derive a condition number estimate of the method. Another space-time Chebyshev collocation scheme which is easier to implement is proposed and analyzed. We also present space-time spectral collocation methods for the Schrodinger, wave, Airy and beam equations. In particular, fully spectral convergence and a condition number estimate are shown for Schrodinger and wave equations. Numerical results verify the theoretical results, and demonstrate that the space-time methods also work for some common nonlinear PDEs (Allen–Cahn, viscous Burgers', Sine–Gordon, nonlinear diffusion, KdV, Kuramoto–Sivashinsky and Cahn–Hilliard equations).

Contribution of authors

The work of this thesis is based on the following publications:

Chapter 2 is a version of a journal article co-authored with Dr. Shaun Lui which has been submitted for publication and is under review: *Superlinear convergence of the methods of Broyden and BFGS based on assumptions about the initial point.*

Chapter 3 is a version of a journal article which has been submitted for publication and is under review. This article is co-authored by Dr. Shaun Lui. I am the primary author of the submitted article: *Superlinear convergence of nonlinear conjugate gradient method and scaled memoryless BFGS method based on assumptions about the initial point.*

Chapter 4 is a version of a journal article co-authored with Dr. Shaun Lui which has been submitted for publication and is under review: *Chebyshev spectral collocation in space and time for PDEs.*

Chapter 5 is part of a manuscript co-authored with Dr. Shaun Lui which is in preparation: *Spectral collocation in space and time for linear PDEs.*

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Shaun Lui. I will be forever grateful for his patience, understanding, expertise and guidance through this thesis.

I would like to thank the staff and faculty in the Department of Mathematics at the University of Manitoba. I also would like to acknowledge for the financial support provided by the Department of Mathematics and the Faculty of Graduate Studies at the University of Manitoba, as well as the Natural Sciences and Engineering Research Council of Canada.

I would like to thank my thesis committee: Dr. Ruppa (Tulsi) Thulasiram, Dr. Craig Cowan and Dr. Martin Gander for their time, and also for the valuable comments that helped me to further improve my thesis.

And the last but not the least, I would like to express my thanks to all my family and friends for their support and encouragement.

This thesis is dedicated to my beloved parents and my dear brothers and sisters for their endless love and continuous support.

Contents

List of Figures

1

Introduction

In first part of this thesis, consisting of Chapters 2 and 3, superlinear convergence of quasi-Newton methods based on assumptions about the initial point is considered. It is well known that the classical Newton's method to solve a nonlinear system of equations converges quadratically if the initial guess is close enough to a solution. One drawback of this theory is that the solution is unknown apriori. Kantorovich's version of this theory only makes assumptions about the initial point and the existence of a solution and the rate of convergence are consequences of the theory [15].

Another disadvantage of the classical Newton's method is that the Jacobian matrix must be formed at every iteration. In practice, the matrix may not be available analytically or its formation may be very expensive. Quasi-Newton methods are designed so that it is relatively inexpensive to compute an approximation to the Jacobian matrix at every iteration. The first and most important contribution is due to Broyden [10], where the matrix approximation from one iteration to the next one can be calculated by a rank-one update. Assuming existence of a root, local convergence of the basic method as well as global convergence of a version with line search are known. See, for instance [24] or [52].

This thesis also addresses the problem of unconstrained minimization of a smooth function $f : \mathbb{R}^N \to \mathbb{R}$. The most popular class of methods for small or medium value

of N is the BFGS method and its variations. While the approximation of the Broyden's method applied to solve the nonlinear system $\nabla f = 0$ is, in general, non-symmetric, the corresponding approximation of the BFGS method is symmetric. Again, local convergence of the basic method and global convergence of a version with line search have been shown [57], [22].

On the other hand, the classical conjugate gradient method was designed to solve a system of linear equations with a symmetric positive definite (SPD) matrix, or equivalently, to find the minimizer of a quadratic objective function with a SPD Hessian. Many variations of the method have been proposed to solve the minimization problem for a general nonlinear function. See [52], for instance. Nonlinear conjugate gradient (NCG) methods are particularly attractive for high-dimensional problems because the memory requirement of the algorithms are $O(N)$. We study the symmetric scaled Perry NCG method ([56]) primarily because, under appropriate assumptions ([77]), it can be considered as a quasi-Newton method with a SPD approximate Hessian (a rank-2 perturbation of the identity) at every iteration.

The method of BFGS approximates the Hessian at every iteration, freeing the user from defining the (exact) Hessian. Unfortunately, the method requires $O(N^2)$ storage and is not feasible for large-scale problems. The memoryless BFGS ([52]) is the BFGS method except that the approximate Hessian is a rank-2 pertubation of the identity matrix. Thus its storage requirement is $O(N)$ and it is attractive for large problems. It is memoryless in the sense that the previous approximation of the Hessian is replaced by the identity. Scaled memoryless BFGS methods ([2]) are those where the approximate Hessian is replaced by a scalar multiple of the identity to reduce the condition number of the approximate Hessian.

The main thrust of the first part of this thesis is to give superlinear local convergence of the methods of Chord, Broyden, BFGS, memoryless BFGS as well as Perry NCG method where all assumptions are made in some region about the initial iterate and hence are verifiable. We shall refer to this as **Kantorovich-type assumptions**. We show existence

of a root or minimizer and superlinear local convergence of these methods without using line search and assuming only conditions about a neighbourhood of the initial point. Following [15], we try to construct a convergence theory with as few constants as possible.

In the second part of this thesis, consisting of Chapters 4 and 5, space-time spectral methods for solving time dependent PDEs are considered. Spectral methods have been used successfully to solve elliptic PDEs for many decades. If the solution is analytic, the numerical solution converges exponentially as a function of the number of spectral modes. For time dependent PDEs, the most common approach is to use low-order finite difference approximation of the time derivative and spectral approximation of the spatial derivatives. This is not ideal since the time discretization error overwhelms the spatial discretization error.

In [49], a Legendre spectral collocation method in both space and time based on the work of Tang and Xu [71] was proposed for the heat equation. The method was shown to converge spectrally when the solution is analytic. A condition number estimate of $O(N^4)$ was derived, where N is the number of spectral modes in each direction. A second space-time method, which is easier to implement and has similar performance was also proposed and analyzed.

The main purpose of second part of this thesis is to demonstrate spectral convergence and $O(N^4)$ condition number estimate for a Chebyshev spectral collocation method. Although much of the basic framework of the theory for the methods based on the two different orthogonal polynomials are similar, the analysis for the Chebyshev case is much more difficult because of the presence of a singular weight function. In this work, a simplified eigenvalue analysis paves the way for a condition number estimate of the Chebyshev space-time method and a similar analysis for other canonical linear PDEs.

In the remainder of this introductory chapter, an outline of the thesis is given. In Chapter 2 we give a simple local convergence theory for the Chord's method for a system of nonlinear equations using Kantorovich-type hypotheses. This is followed by local super-

linear convergence of Broyden's and BFGS method using Kantorovich-type assumptions. For the latter, we introduce a norm which depends on the iteration number to estimate the difference between inverses of the approximate and exact Jacobians. This idea may be applicable in other situations.

In Chapter 3, local superlinear convergence of the symmetric scaled Perry NCG method is given using Kantorovich-type assumptions. This is followed by an analogous theory for a generalized scaled memoryless BFGS method.

In the Chapter 4, the space-time spectral convergence of the Chebyshev spectral collocation method of Tang and Xu for the 1D heat equation is established. The condition number of the method is shown to be $O(N^4)$. A similar space-time spectral collocation method which is easier to implement for more general PDEs and which exhibits nearly identical characteristics is proposed and analyzed. Some simple iterative schemes for two nonlinear PDEs (Allen–Cahn and viscous Burgers' equations) are briefly discussed and some numerical experiments in MATLAB are shown to confirm the theoretical results.

In the Chapter 5, a space-time Chebyshev spectral collocation method for the 1D Schrodinger, wave, Airy and beam equations are introduced. A condition number estimate of the method for Schrodinger and wave equations is shown. Basically, the condition number is bounded by a multiple of the condition number of the spectral approximation of the associated spatial differential operator. Also some simple iterative schemes for four nonlinear PDEs (Sine–Gordon, nonlinear diffusion, KdV, Kuramoto–Sivashinsky and Cahn–Hilliard equations) are briefly discussed. Numerical experiments in MATLAB confirm the theoretical results.

Finally in Chapter 6, we summarize and offer some open problems and directions of future work.

2

Quasi-Newton methods, Chord's, Broyden's and BFGS methods

In this chapter we establish superlinear convergence of a class of quasi-Newton methods without applying line search and by using Kantorovich-type hypotheses. Existence of a solution is a consequence of the theory.

2.1 Introduction

Let Ω be an open set in $\mathbb{R}^N$. Given a smooth $F : \Omega \to \mathbb{R}^N$, the problem of interest is to find $x^* \in \Omega$ so that $F(x^*) = 0$. A classical method to solve this problem is Newton's method. Given an initial guess $x_0 \in \Omega$ for the root of the function, Newton's method produces a sequence of iterates $\{x_n\}$ defined by:

$$x_{n+1} = x_n - F'(x_n)^{-1}F(x_n), \qquad n \geq 0,$$

where $F'(x_n)$ is the Jacobian matrix evaluated at x_n. Any method that replaces the exact Jacobian $F'(x_n)$ with an approximation is a quasi-Newton method. Most practical

algorithms for finding a zero of F are given by

$$x_{n+1} = x_n - A_n^{-1} F(x_n), \qquad n \geq 0,$$

where A_n is a sequence of nonsingular matrices. The main work here is to derive a convergence theory for quasi-Newton methods where all assumptions can be verified and existence of a root is a consequence of the theory. Also, the theory should have as few parameters as possible. Before deriving the general convergence theorems for these methods, we will need to discuss rates of convergence.

Definition 2.1. Assume $\{x_n\} \subseteq \mathbb{R}^N$ converges to x^*. $\{x_n\}$ converges superlinearly to x^* if and only if either $x_n = x^*$ for all sufficiently large n or $x_n \neq x^*$ for $n \geq n_0$ and

$$\lim_{n\to\infty} \frac{\|x_{n+1} - x^*\|}{\|x_n - x^*\|} = 0.$$

Throughout this chapter and next, let $\|\cdot\|$ denote the Euclidean vector or matrix norm and $B_r(x)$ denote the open ball of radius r with center at x. Recall that for any $N \times N$ matrix A, $\|A\|_F^2 = \sum_{i,j} a_{ij}^2$, and $\|A\| \leq \|A\|_F \leq \sqrt{N}\|A\|$. The following theorem gives simple assumptions for quadratic convergence of Newton's method.

Theorem 2.2 (Dennis and Schnabel, 1996, [24])**.** Suppose $F : \mathbb{R}^N \to \mathbb{R}^N$ is C^1. Assume $F(x^*) = 0$ for some $x^* \in \mathbb{R}^N$. Suppose $F'(x^*)$ is invertible and there are some positive constants α, β and r such that $\|F'(x)^{-1}\| \leq \alpha$ for all $x \in \overline{B_r(x^*)}$, $\|F'(x) - F'(y)\| \leq \beta\|x - y\|$ for all $x, y \in \overline{B_r(x^*)}$, and $\alpha\beta r < 2$. If $\|x_0 - x^*\| \leq r$, then the Newton's iteration is well defined, convergent to x^* and

$$\|x_{n+1} - x^*\| \leq \alpha\beta \frac{\|x_n - x^*\|^2}{2}, \qquad \forall n \geq 0.$$

The above theorem assumes the existence of a solution and a bound on the initial error, which are both unknown, in general. Another convergence result for Newton's method was

introduced by L. Kantorovich [41]. It makes no assumption about the existence of a root. It proves that if $F'(x_0)$ is nonsingular, F' is Lipschitz continuous in a region containing x_0, and the first step of Newton's method is sufficiently small, then there must be a root in this region, and also it is unique.

Theorem 2.3 (Ciarlet, 2012, [16])**.** Let Ω be an open convex subset in $\mathbb{R}^N$ and $x_0 \in \Omega$. Suppose $F : \Omega \to \mathbb{R}^N$ is $C^1(\Omega)$ and $F'(x_0)$ is invertible. Assume there is some $r > 0$ satisfying three hypotheses:

i) $\overline{B_r(x_0)} \subset \Omega$,

ii) $\|F'(x_0)^{-1}F(x_0)\| \leq r/2$,

iii) $\|F'(x_0)^{-1}(F'(y) - F'(x))\| \leq \|y - x\|/r$ for all $x, y \in B_r(x_0)$.

Then

i) $F'(x)$ is invertible for each $x \in B_r(x_0)$,

ii) every Newton's iterate $x_n \in B_r(x_0), n \geq 0$,

iii) $x_n \to x^* \in \overline{B_r(x_0)}$, where $F(x^*) = 0$,

iv) $\|x_n - x^*\| \leq 2^{-n}r$ for all $n \geq 0$,

v) x^* is the only zero of F in $\overline{B_r(x_0)}$.

Note that all assumptions of the above theorem are about x_0 and so verifiable. Furthermore, the existence of a solution is a consequence of the theorem. So the above theorem can be used to demonstrate that a given nonlinear system has a solution.

While Newton's method converges quadratically near the solution, it requires the formation of the Jacobian which can sometimes be expensive and impractical. Broyden in [10] suggested an algorithm for finding a solution of a system of nonlinear equations where no Jacobian information is needed. For a function $F : \mathbb{R}^N \to \mathbb{R}^N$, the Broyden's method is defined by

$$x_{n+1} = x_n - A_n^{-1}F(x_n), \quad n \geq 0,$$

where

$$A_{n+1} = A_n + \frac{F(x_{n+1})s_n^T}{\|s_n\|^2},$$

$s_n = x_{n+1} - x_n$ and A_0 is a given, invertible matrix and $x_0 \in \Omega$ is an initial guess. In practice, A_0 can be taken as I or $F'(x_0)$ if it is invertible. In [11] it is shown that the iterates converge superlinearly provided that x_0 and A_0 are sufficiently close to the quantities that they approximate.

Theorem 2.4 (Broyden, Dennis and Moré, 1973, [11])**.** Let F be differentiable in $B_r(x^*)$, an open ball about a root x^*. Assume that there are some positive constants β and γ so that $\|F'(x^*)^{-1}\| \leq \beta$ and $\|F'(x) - F'(y)\| \leq \gamma\|x - y\|$ for all $x, y \in B$. Suppose $x_0 \in B$ and A_0 satisfies

$$\|A_0 - F'(x_0)\| + 2\gamma\|x_0 - x^*\| \leq (8\beta)^{-1}.$$

Then the iterates $\{x_n\}$ given by Broyden's method are well defined and converge superlinearly to x^*.

We remark that A_n does not converge to $F'(x^*)$ in general.

Notice that almost all works in convergence theory of Broyden's method have been done by using assumptions of existence of a solution and a bound on $\|x^* - x_0\|$ and $\|F'(x^*) - A_0\|$. Dennis in [21] has proposed a Kantorovich-type analysis (assumptions based only on the initial guess x_0 and A_0) for Broyden's method. With $F'(x_0)$ nonsingular, F Lipschitz continuous in a region containing x_0, and assuming three constants δ, β and η, such that

$$\|F'(x_0) - A_0\| \leq \delta, \quad \|A_0^{-1}\| \leq \beta, \quad \|A_0^{-1}F(x_0)\| \leq \eta,$$

local convergence of this method has been proved but without superlinear convergence. In our work, we reduce the number of constants to one and demonstrate superlinear convergence.

In optimization, quasi-Newton methods can be applied for finding local minima of multi-variable functions. Suppose we want to find a local minimum point $x^* \in \mathbb{R}^N$ of $f : \mathbb{R}^N \to \mathbb{R}$ so that $\nabla f(x^*) = 0$ and Hessian matrix $D^2 f(x^*)$ is an $N \times N$ SPD matrix. Using Newton's method:

$$x_{n+1} = x_n - D^2 f(x_n)^{-1} \nabla f(x_n), \quad n \geq 0,$$

this iteration converges quadratically provided that x_0 is sufficiently close to x^*. When N is large, Newton's method may not be efficient because of the need to form the large Jacobian matrix at every iteration. One possibility is to use Broyden's method. Unfortunately, the approximate Jacobian in Broyden's method is, in general, non-symmetric, in contrast to the symmetric $D^2 f(x)$. Assume $x_0 \in \Omega$, A_0 is a SPD matrix. The BFGS iteration for finding a local minimum is given by

$$x_{n+1} = x_n - A_n^{-1} \nabla f(x_n), \quad n \geq 0,$$

where a rank-two Jacobian update is given by

$$A_{n+1} = A_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{A_n s_n s_n^T A_n}{s_n^T A_n s_n},$$

for any $n \geq 0$, where $s_n = x_{n+1} - x_n$ and $y_n = F(x_{n+1}) - F(x_n)$. If $y_n^T s_n > 0$ in each step, the BFGS approximate Jacobian A_{n+1} stays SPD [52].

BFGS method with line search is globally convergent for convex functions. The analysis is based on early work by Powel [57] and Dennis and Moré [22]. Suppose that the starting point is sufficiently close to the solution x^* and that the initial Hessian approximation is sufficiently close to $F'(x^*)$. Then BFGS with line search converges superlinearly.

Theorem 2.5 (Nocedal and Wright, [52], Chapter 6)**.** Let Ω be an open convex set in $\mathbb{R}^N$ and $f : \Omega \to \mathbb{R}$ be twice continuously differentiable. Let $x^* \in \Omega$ so that $\nabla f(x^*) = 0$ and

$D^2 f(x^*)$ is SPD. Suppose there is some positive constant m so that it is a lower bound for all eigenvalues of $D^2 f(x)$ for all $x \in \Omega$. Assume that

$$\|D^2 f(x) - D^2 f(y)\| \leq L\|x - y\|,$$

for some positive constant L and all $x, y \in \Omega$. Let $\{x_n\}$ be the sequence produced by the BFGS method with line search and $x_n \neq x^*$ for all n. Then $x_n \to x^*$ superlinearly.

Again the main assumption is existence of a solution x^*, which is an unknown quantity in general. Line search enables the algorithm to be globally convergent.

The aim here is to show superlinear convergence of Chord's, Broyden's and BFGS methods without applying line search and by using Kantorovich-type hypotheses. Part of the attraction of this theory is that the number of constants has been reduced, in the spirit of [16].

In the remainder of this chapter, in Section 2.2, some lemmas are given that are necessary in the proof of theorems and propositions. In Section 2.3 local superlinear convergence of Chord's method is given using Kantorovich-type assumptions. This is followed by an analogous theory for superlinear convergence of Broyden's and BFGS method in Sections 2.4 and 2.5.

2.2 Preliminaries

The following lemmas are needed in the proof of theorems coming in the next sections.

Lemma 2.6. Let A, B be SPD, then

$$\|AB\|^2 \leq \|A^2\, B^2\|.$$

Proof. Define inner product $\langle x, y\rangle = x^T A^{-1}\, y$. It is well known that AB is self adjoint with respect to this inner product and is positive definite. Let $\lambda_{max}(M)$ be the maximum

eigenvalue of matrix M, then

$$\lambda_{max}(AB) = \max_{y \neq 0} \frac{\langle ABy, y \rangle}{\langle y, y \rangle} = \max_{y \neq 0} \frac{y^T By}{y^T A^{-1} y}.$$

Since $\lambda_{max}(A^2B^2) \leq \|A^2B^2\|$, it follows that

$$\|AB\|^2 = \|AB(AB)^T\| = \|AB^2A\| = \max_{x \neq 0} \frac{x^T AB^2 A\, x}{x^T x} = \max_{y \neq 0} \frac{y^T B^2\, y}{y^T A^{-2} y} = \lambda_{max}(A^2B^2) \leq \|A^2B^2\|.$$

□

Lemma 2.7 (Sherman and Morrison, 1949, [23], Lemma 4.2)**.** Let $u, v \in \mathbb{R}^N$ and assume $A \in \mathbb{R}^{N \times N}$ is nonsingular. Then $A + uv^T$ is nonsingular if and only if $\sigma = 1 + v^T A^{-1} u \neq 0$. If $\sigma \neq 0$, then

$$(A + uv^T)^{-1} = A^{-1} - (1/\sigma) A^{-1} uv^T A^{-1}.$$

Lemma 2.8 (Dennis and Moré, 1977, [23], Lemma 8.5)**.** Let u, v be vectors so that $u^T v \neq 0$. Then

$$\left\| I - \frac{uv^T}{v^T u} \right\| = \frac{\|u\| \|v\|}{|v^T u|}.$$

Lemma 2.9 (Dennis and Schnabel, 1996, [24], Theorem 3.1.4)**.** Let A be a square matrix and $\|I - A\| < 1$. Then A is invertible and

$$\|A^{-1}\| \leq \frac{1}{1 - \|I - A\|}.$$

Lemma 2.10 (Dennis and Moré, 1996, [23], Lemma 3.2)**.** Let u, v be non-zero vectors so that $\|u - v\| \leq \lambda \|u\|$ for some $\lambda \in (0, 1)$. Then

$$1 - \left(\frac{u^T v}{\|u\| \|v\|} \right)^2 \leq \lambda^2.$$

2.3 Chord's method

The Chord's method to solve the nonlinear system $F(x) = 0$ is given by the iteration

$$x_{n+1} = x_n - A^{-1}F(x_n), \qquad n \geq 0,$$

where x_0 is an initial guess and $A = F'(x_0)$ is invertible. The Chord's method is Newton's iteration except that the Jacobian is fixed at A for all n. This is an alternative to Newton's method because the Jacobian is formed only once in the beginning. The drawback is that the convergence is only linear. Below is a local convergence theory using Kantorovich-type assumptions.

Theorem 2.11. Let Ω be an open set in $\mathbb{R}^N$ and $F : \Omega \to \mathbb{R}^N$ be continuously differentiable on Ω. Given $x_0 \in \Omega$. Suppose $A = F'(x_0)$ is non-singular. Assume for some $r \in (0, 1)$ that $\overline{B_r(x_0)} \subset \Omega$, $\|A^{-1}F(x_0)\| \leq (1 - r)r$ and

$$\|A^{-1}(F'(v) - F'(w))\| \leq \|v - w\|, \qquad v, w \in B_r(x_0).$$

Let $\{x_n\}$ be the iterates of the Chord's method. Then $x_n \to x^* \in \overline{B_r(x_0)}$, where $F(x^*) = 0$. Let $e_n = x_n - x^*$. Then $\|e_n\| \leq r^{n+1}$, $n \geq 0$. Furthermore, x^* is the unique zero of F in $\overline{B_r(x_0)}$.

Proof. Let $s_n = x_{n+1} - x_n$. We claim by induction that $x_{n+1} \in B_r(x_0)$ and $\|s_n\| \leq (1 - r)r^{n+1}$, $\forall n \geq 0$.

The base case $n = 0$ holds trivially since $s_0 = -A^{-1}F(x_0)$ and so by hypothesis, $\|s_0\| \leq (1 - r)r < r$. This also shows that $x_1 \in B_r(x_0)$. Assume that the claims hold for $n - 1$. We show that they also hold for n.

By the Mean Value Theorem, there is some ξ along the line joining x_n and x_{n-1} so

that $F(x_n) - F(x_{n-1}) = F'(\xi)s_{n-1}$. By the induction hypothesis,

$$\begin{aligned}
\|s_n\| &= \|A^{-1}F(x_n)\| = \|A^{-1}\Big(F(x_n) - F(x_{n-1})\Big) - s_{n-1}\| \\
&\leq \|\Big(A^{-1}F'(\xi) - I\Big)s_{n-1}\| = \|A^{-1}\Big(F'(\xi) - F'(x_0)\Big)s_{n-1}\| \\
&\leq \|\xi - x_0\|\,(1-r)r^n \leq (1-r)r^{n+1}.
\end{aligned}$$

Since $x_{n+1} - x_0 = \sum_{j=0}^{n} s_j$, it follows that $\|x_{n+1} - x_0\| \leq \sum_{j=0}^{n}(1-r)r^{j+1} < r$, or $x_{n+1} \in B_r(x_0)$. For any non-negative p, we have $x_{n+p+1} - x_n = \sum_{j=n}^{n+p} s_j$, and so

$$\|x_{n+p+1} - x_n\| \leq (1-r)\sum_{j=n}^{n+p} r^{j+1} \leq r^{n+1}.$$

This implies that $\{x_n\}$ is a Cauchy sequence and so it must converge to some $x^* \in \overline{B_r(x_0)}$. Also, taking $p \to \infty$,

$$\|e_n\| \leq r^{n+1}.$$

Consequently, $A^{-1}F(x_n) = -s_n \to 0$. This shows that $F(x^*) = 0$.

Let $\hat{x}$ be any zero of F in $\overline{B_r(x_0)}$. Define $\hat{e}_n = x_n - \hat{x}$. We show $\|\hat{e}_n\| \leq r^{n+1}$ by induction. The base case is trivial. Suppose the claim is true for n. There is some ξ in between x_n and $\hat{x}$ so that $F(x_n) - F(\hat{x}) = F'(\xi)(x_n - \hat{x})$. Then

$$\hat{e}_{n+1} = x_{n+1} - \hat{x} = x_n - A^{-1}F(x_n) + A^{-1}F(\hat{x}) - \hat{x} = \hat{e}_n - A^{-1}F'(\xi)\hat{e}_n.$$

Therefore

$$\begin{aligned}
\|\hat{e}_{n+1}\| &\leq \|A^{-1}\Big(F'(x_0) - F'(\xi)\Big)\|\;\|\hat{e}_n\| \\
&\leq \|x_0 - \xi\|\; r^{n+1} \leq r^{n+2}.
\end{aligned}$$

As a result,

$$\|x^* - \hat{x}\| \leq \|x^* - x_n\| + \|x_n - \hat{x}\| \leq 2r^{n+1} \to 0. \tag{2.1}$$

Hence $x^* = \hat{x}$. □

2.4 Broyden's method

In Newton's method we need to form the Jacobian $F'(x_n)$ at every iteration, which may be computationally intensive, or may not be available analytically. Broyden in [10] devised an approximate Jacobian which can calculated from the approximate Jacobian of the previous iteration by a rank-one update. Given $x_0 \in \Omega$ and an invertible initial approximate Jacobian A_0 the algorithm is

$$\begin{aligned} x_{n+1} &= x_n + s_n, \qquad s_n = -A_n^{-1} F(x_n), \qquad n \geq 0, \\ A_{n+1} &= A_n + \frac{F(x_{n+1}) s_n^T}{\|s_n\|^2} \ . \end{aligned}$$

Using classical assumptions (existence of a solution x^* and bounds on F and F' in a neighbourhood of x^*), local superlinear convergence and global convergence of the method with line search are known. See, for instance, [24] or [52].

Since x^* is not known apriori, the assumptions cannot be checked in practice. The purpose of this section is to give a local superlinear convergence of Broyden's method using Kantorovich-type assumptions.

We are now ready to show local convergence of Broyden's method, which will be followed by a proof of superlinear convergence. Our technique of proof combines the elegant Newton-Kantorovich theory with only one constant (Theorem 7.7.5 in [15]) and the local convergence of Broyden's method ([24]). Note that [21] has shown a local Kantorovich-type convergence result, but without superlinear convergence.

Theorem 2.12. Let Ω be open in $\mathbb{R}^N$, $F : \Omega \to \mathbb{R}^N$, $F \in C^1(\Omega)$, $x_0 \in \Omega$ and A_0 invertible. For some $0 < r \leq 1/2$ assume $\overline{B_r(x_0)} \subset \Omega$ and

$$\|F'(x_0)^{-1}F(x_0)\| \leq \xi r^2, \tag{2.2}$$

$$\|F'(x_0)^{-1}(F'(u) - F'(v))\| \leq \frac{\eta\|u - v\|}{r}, \qquad \forall u, v \in B_r(x_0), \tag{2.3}$$

$$\|I - F'(x_0)^{-1}A_0\| \leq dr, \tag{2.4}$$

where ξ, η and d are positive constants dependent on r (to be defined later). Then Broyden's iteration $\{x_n\}$ is well defined and exactly one of the following cases holds,
(i) $F(x_n) = 0$ for some $n \geq 0$.
(ii) Broyden's method converges to a unique zero of F in $\overline{B_r(x_0)}$.

Proof. Define $G(y) = F'(x_0)^{-1}F(y)$. By this definition, $F(x^*) = 0$ if and only if $G(x^*) = 0$, zeros of F are zeros of G and also $G'(y) = F'(x_0)^{-1}F'(y)$, G is differentiable as F is. Define

$$\begin{aligned} &B_0 = F'(x_0)^{-1}A_0, \quad y_0 = x_0, \\ &y_{n+1} = y_n + t_n, \qquad t_n = -B_n^{-1}G(y_n), \qquad n \geq 0, \\ &B_{n+1} = B_n + \frac{G(y_{n+1})t_n^T}{\|t_n\|^2} \ . \end{aligned}$$

Assume $F(x_n) \neq 0$ for all $n \geq 0$. First we show that $\forall n \geq 0$, $y_n = x_n$ and $B_n = F'(x_0)^{-1}A_n$. We use mathematical induction for proving these statements. Basic step is true obviously, $y_0 = x_0$ and $B_0 = F'(x_0)^{-1}A_0$, by using definition. Let $x_n = y_n$ and $B_n = F'(x_0)^{-1}A_n$, for some positive integer n, then we need to show $x_{n+1} = y_{n+1}$ and $B_{n+1} = F'(x_0)^{-1}A_{n+1}$. Notice that:

$$\begin{aligned} t_n &= -B_n^{-1}G(x_n) = -(F'(x_0)^{-1}A_n)^{-1}F'(x_0)^{-1}F(x_n) \\ &= -A_n^{-1}F'(x_0)F'(x_0)^{-1}F(x_n) = -A_n^{-1}F(x_n) = s_n, \end{aligned}$$

therefore $t_n = s_n$ and also $y_{n+1} = y_n + t_n = x_n + s_n = x_{n+1}$. By definition of B_{n+1} we get:

$$\begin{aligned} B_{n+1} &= B_n + \frac{G(x_{n+1})t_n^T}{\|t_n\|^2} = F'(x_0)^{-1}A_n + \frac{F'(x_0)^{-1}F(x_{n+1})s_n^T}{\|s_n\|^2} \\ &= F'(x_0)^{-1}(A_n + \frac{F(x_{n+1})s_n^T}{\|s_n\|^2}) = F'(x_0)^{-1}A_{n+1}. \end{aligned}$$

Furthermore, it is easy to show that by using assumptions of the theorem,

$$\begin{aligned} &\|G(x_0)\| \leq \xi r^2; \\ &\|G'(u) - G'(v)\| \leq \frac{\eta\|u-v\|}{r}, \qquad \forall u, v \in B_r(x_0)\,. \end{aligned}$$

Also two following claims are really useful.

Claim 1. $\|G'(u)^{-1}\| \leq \dfrac{1}{1-\eta\|x_0-u\|/r}$, for all $u \in B_r(x_0)$.

We have $G'(x_0) = F'(x_0)^{-1}F'(x_0) = I$, therefore for $u \in B_r(x_0)$,

$$\|I - G'(u)\| = \|G'(x_0) - G'(u)\| \leq \frac{\eta\|x_0-u\|}{r} < \eta.$$

If we assume $\eta < 1$, then by using Lemma 2.9, $G'(u)$ is invertible and

$$\|G'(u)^{-1}\| \leq \frac{1}{1-\|I-G'(u)\|} \leq \frac{1}{1-\eta\|x_0-u\|/r}\,.$$

Claim 2. $\|G(u) - G(v) - G'(v)(u-v)\| \leq \dfrac{\eta\|u-v\|^2}{2r}$, for all $u, v \in B_r(x_0)$.

$$\begin{aligned} \|G(u) - G(v) - G'(v)(u-v)\| &= \left\| \int_0^1 \Big(G'(tu + (1-t)v) - G'(v) \Big)(u-v)dt \right\| \\ &\leq \|u-v\| \int_0^1 \|G'\Big(\ tu + (1-t)v\ \Big) - G'(v)\|dt \\ &\leq \frac{\eta\|u-v\|}{r} \int_0^1 \|t(u-v)\|dt \\ &= \frac{\eta\|u-v\|^2}{r} \int_0^1 t\ \ dt = \frac{\eta\|u-v\|^2}{2r}. \end{aligned}$$

Claim: There are some positive constants α, μ and β dependent on r (to be defined later), such that for $n \geq 0$,

3. $\|x_n - x_0\| \leq r(1 - r^n)$;

4. $\|G(x_n)\| \leq \xi\ r^{n+2}$;

5. $\|G'(x_n) - B_n\| \leq \alpha r$;

6. $G'(x_n)$ is invertible and $\|G'(x_n)^{-1}\| \leq \mu$;

7. B_n is invertible and $\|B_n^{-1}\| \leq \beta$;

8. $\|s_n\| \leq r^{n+2}$.

In following the proof of these claims is given by using mathematical induction. Basic step for Claim 3 is trivial. By definition of B_0, it is invertible thus s_0 is well defined and x^1 exists. Also $\|G'(x_0) - B_0\| = \|I - F'(x_0)^{-1}A_0\| \leq dr \leq \alpha r$, if we choose α such that $d \leq \alpha$. By assumption $\|G'(x_0)^{-1}\| = 1$, let $\mu \geq 1$, so we get: $\|G'(x_0)^{-1}\| \leq \mu$. In addition $\|I - B_0\| = \|I - F'(x_0)^{-1}A_0\| \leq dr$. If we assume $dr < 1$, then $\|I - B_0\| < 1$ and by using Lemma 2.9

$$\|B_0^{-1}\| \leq \frac{1}{1 - \|I - B_0\|} \leq \frac{1}{1 - dr}.$$

So by assuming $\beta \geq \dfrac{1}{1 - dr}$, we have $\|B_0^{-1}\| \leq \beta$. Also $\|s_0\| = \| - B_0^{-1}G(x_0)\| \leq \|B_0^{-1}\|\|G(x_0)\| \leq \beta\xi r^2 \leq\ r^2$, by assuming $\beta\xi \leq 1$. Then we assume all of the statements are true for some integer $n \geq 1$, we will show they hold for $n + 1$.

Claim 3. Since B_n is invertible by hypothesis of induction, x_{n+1} exists and

$$\|x_{n+1} - x_0\| \leq \|x_{n+1} - x_n\| + \|x_n - x_0\| \leq\ r^{n+2} + r(1 - r^n) = r(1 - r^n(1 - r)) \leq r(1 - r^{n+1}),$$

since $r \leq 1/2$. $\|x_{n+1} - x_0\| \leq r(1 - r^{n+1}) \leq r$, and also $x_{n+1} \in B_r(x_0)$.

Claim 4. We have $s_n = -B_n^{-1}G(x_n)$, so $G(x_n) = -B_n s_n$. Then

$$\begin{aligned}
\|G(x_{n+1})\| &= \|G(x_{n+1}) - G(x_n) + G(x_n)\| = \|G(x_{n+1}) - G(x_n) - B_n s_n\| \\
&= \|G(x_{n+1}) - G(x_n) - G'(x_n)s_n + (G'(x_n) - B_n)s_n\| \\
&\leq \|G(x_{n+1}) - G(x_n) - G'(x_n)s_n\| + \|(G'(x_n) - B_n)s_n\| \\
&\leq \frac{\eta\|s_n\|^2}{2r} + \alpha r\|s_n\| = \|s_n\|(\frac{\eta\|s_n\|}{2r} + \alpha r) \qquad (2.5) \\
&\leq r^{n+2}\,(\frac{\eta r^{n+2}}{2r} + \alpha r) \\
&\leq r^{n+3}\,(\eta + \alpha) \leq \xi r^{n+3},
\end{aligned}$$

if we assume $\eta + \alpha \leq \xi$. Then $\|G(x_{n+1})\| \leq \xi\ r^{n+3}$, as we need.

Claim 5. Observe that

$$\begin{aligned}
\|G'(x_{n+1}) - B_{n+1}\| &= \left\|G'(x_{n+1}) + G'(x_n) - G'(x_n) - B_n - \frac{G(x_{n+1})s_n^T}{\|s_n\|^2}\right\| \\
&\leq \|G'(x_{n+1}) - G'(x_n)\| + \left\|G'(x_n) - B_n - \frac{G(x_{n+1})s_n^T}{\|s_n\|^2}\right\|. \qquad (2.6)
\end{aligned}$$

Consider the second term of this inequality, we could write it as:

$$\begin{aligned}
G'(x_n) - B_n - \frac{G(x_{n+1})s_n^T}{\|s_n\|^2} &= G'(x_n) - B_n - \frac{\Big(G(x_{n+1}) - G(x_n) + G(x_n)\Big)s_n^T}{\|s_n\|^2} \\
&= G'(x_n) - B_n - \frac{\Big(G(x_{n+1}) - G(x_n)\Big)s_n^T}{\|s_n\|^2} - \frac{G(x_n)s_n^T}{\|s_n\|^2} \\
&= G'(x_n) - B_n - \frac{\Big(G(x_{n+1}) - G(x_n)\Big)s_n^T}{\|s_n\|^2} + \frac{B_n s_n s_n^T}{\|s_n\|^2} \\
&= G'(x_n) - B_n - \int_0^1 G'(\ (1-t)x_n + tx_{n+1}\ )\frac{s_n s_n^T}{\|s_n\|^2}\ dt + \frac{B_n s_n s_n^T}{\|s_n\|^2} \\
&= G'(x_n) - B_n + \int_0^1 \left(G'(x_n) - G'(\ (1-t)x_n + tx_{n+1}\ )\right)\frac{s_n s_n^T}{\|s_n\|^2}\ dt \\
&\qquad - \int_0^1 G'(x_n)\frac{s_n s_n^T}{\|s_n\|^2}\ dt + \frac{B_n s_n s_n^T}{\|s_n\|^2} \\
&= (G'(x_n) - B_n)(I - \frac{s_n s_n^T}{\|s_n\|^2}) \\
&\qquad + \int_0^1 [G'(x_n) - G'((1-t)x_n + tx_{n+1})]\frac{s_n s_n^T}{\|s_n\|^2}\ dt.
\end{aligned}$$

Therefore:

$$\begin{aligned}
\left\|G'(x_n) - B_n - \frac{G(x_{n+1})s_n^T}{\|s_n\|^2}\right\| &\le \|G'(x_n) - B_n\|\left\|I - \frac{s_n s_n^T}{\|s_n\|^2}\right\| \\
&\qquad + \int_0^1 \|G'(x_n) - G'(\ (1-t)x_n + tx_{n+1}\ )\|\frac{\|s_n\|\|s_n^T\|}{\|s_n\|^2}\ dt \\
&\le \|G'(x_n) - B_n\| + \int_0^1 \frac{\eta\|x_n - [(1-t)x_n + tx_{n+1}]\|}{r}\ dt \\
&\le \|G'(x_n) - B_n\| + \int_0^1 \frac{t\eta\|s_n\|}{r}\ dt \\
&\le \|G'(x_n) - B_n\| + \frac{\eta\|s_n\|}{2r}\ .
\end{aligned}$$

Substitute this in inequality (2.6) we get:

$$\begin{aligned}
\|G'(x_{n+1}) - B_{n+1}\| &\leq \|G'(x_{n+1}) - G'(x_n)\| + \|G'(x_n) - B_n\| + \frac{\eta\|s_n\|}{2r} \\
&\leq \frac{\eta\|s_n\|}{r} + \|G'(x_n) - B_n\| + \frac{\eta\|s_n\|}{2r} \\
&= \frac{3\eta\|s_n\|}{2r} + \|G'(x_n) - B_n\| \\
&\leq \frac{3\eta}{2r}(\|s_n\| + \|s_{n-1}\| + \ldots + \|s_0\|\ ) + \|G'(x_0) - B_0\| \\
&\leq \frac{3\eta}{2r}\Big(\ r^{n+2} + r^{n+1} + \ldots + r^2) + dr \\
&\leq \frac{3\eta r}{2}\Big(\frac{1 - r^{n+1}}{1 - r}\Big) + dr \leq 3\eta r + dr \leq \alpha r,
\end{aligned}$$

if we choose α such that $\alpha \geq 3\eta + d$.

Claim 6. By using Claim 3, we have $\|x_{n+1} - x_0\| \leq\ r(1 - r^{n+1}) \leq r$, and also $x_{n+1} \in B_r(x_0)$. Then by using Claim 1, $G'(x_{n+1})$ is invertible and

$$\|G'(x_{n+1})^{-1}\| \leq \frac{1}{1 - \dfrac{\eta\|x_{n+1} - x_0\|}{r}} \leq \frac{1}{1 - \eta}\ .$$

Define $\mu = \dfrac{1}{1 - \eta} > 1$. Then $\|G'(x_{n+1})^{-1}\| \leq \mu$ as we need.

Claim 7. Notice that:

$$G'(x_{n+1})^{-1}B_{n+1} = I + G'(x_{n+1})^{-1}\Big(B_{n+1} - G'(x_{n+1})\ \Big), \tag{2.7}$$

and

$$\|G'(x_{n+1})^{-1}(B_{n+1} - G'(x_{n+1}))\| \leq\ \|G'(x_{n+1})^{-1}\|\ \|B_{n+1} - G'(x_{n+1})\|\ \leq \mu\alpha r.$$

Assume $\mu\alpha r < 1$, then $G'(x_{n+1})^{-1}B_{n+1}$ is invertible which means B_{n+1} is invertible and

$$\begin{aligned}\left\|\left(I + G'(x_{n+1})^{-1}\left(B_n - G'(x_{n+1})\right)\right)^{-1}\right\| &\leq \frac{1}{1 - \|G'(x_{n+1})^{-1}\Big(B_{n+1} - G'(x_{n+1})\Big)\|} \\ &\leq \frac{1}{1-\mu\alpha r}.\end{aligned}$$

From (2.7), $B_{n+1}^{-1} = \left(I + G'(x_{n+1})^{-1}\left(B_n - G'(x_{n+1})\right)\right)^{-1} G'(x_{n+1})^{-1}$,

$$\|B_{n+1}^{-1}\| \leq \left\|\left(I + G'(x_{n+1})^{-1}\left(B_{n+1} - G'(x_{n+1})\right)\right)^{-1}\right\| \left\|G'(x_{n+1})^{-1}\right\| \leq \frac{\mu}{1-\mu\alpha r}.$$

Let $\beta = \max\{\frac{1}{1-dr}, \frac{\mu}{1-\mu\alpha r}\}$, then $\|B_{n+1}^{-1}\| \leq \beta$. Notice that

$$\frac{\mu}{1-\mu\alpha r} = \frac{1}{1-\eta-\alpha r} = 1 + \frac{\eta+\alpha r}{1-\eta-\alpha r} > 1,$$

so $\beta > 1$.

Claim 8. Since B_{n+1} is invertible, s_{n+1} is well defined and

$$\|s_{n+1}\| = \| - B_{n+1}^{-1}G(x_{n+1})\| \leq \|B_{n+1}^{-1}\|\|G(x_{n+1})\| \leq \beta\xi r^{n+3} \leq r^{n+3},$$

since $\beta\xi \leq 1$.

Therefore by using mathematical induction we have the results. By using Claim 8 we could say $\{x_n\}$ is a Cauchy sequence lying in $B_r(x_0)$. Given $p, q \geq 0$ we have:

$$\|x_p - x_{p+q}\| \leq \sum_{k=p}^{p+q-1} \|x_{k+1} - x_k\| \leq \sum_{k=p}^{p+q-1} r^{k+2} < r^2 \sum_{k=p}^{\infty} r^k = \frac{r^{p+2}}{1-r} \leq r^{p+1},$$

since $r \leq \frac{1}{2} \Rightarrow \frac{1}{1-r} \leq \frac{1}{r}$, therefore $\{x_n\}$ converges to a point $x^* \in \overline{B_r(x_0)}$. By using the fact that G is a continuous function and $\|G(x_n)\| \leq \xi r^{n+2}$, it follows that $G(x^*) = 0$, which implies $F(x^*) = 0$. By taking $q \to \infty$ and $p = n$ in the above calculation we get

$\|x_n - x^*\| \leq r^{n+1}$. Let $e_n = x_n - x^*$, then we have $\|e_n\| \leq r^{n+1}$. Observe that $x \in \overline{B_r(x_0)}$. Now for proof of uniqueness, let $\hat{x}$ be any other zero of F in $\overline{B_r(x_0)}$. Then we could show that $\|\hat{e}_{n+1}\| \leq \dfrac{\|\hat{e}_n\|}{2}$ for all $n \geq 0$, where $\hat{e}_n = x_n - \hat{x}$. Notice that:

$$\begin{aligned}
\hat{e}_{n+1} &= x_{n+1} - \hat{x} = x_n + s_n - \hat{x} = x_n - B_n^{-1}G(x_n) - \hat{x} \\
&= B_n^{-1}B_n\,\hat{e}_n + B_n^{-1}\Big(- G(x_n) + G(\hat{x})\Big) \\
&= B_n^{-1}\Big(B_n - G'(x_n)\Big)\,\hat{e}_n + B_n^{-1}\Big(- G(x_n) + G(\hat{x}) + G'(x_n)\hat{e}_n\Big).
\end{aligned}$$

Therefore:

$$\begin{aligned}
\|\hat{e}_{n+1}\| &= \left\|B_n^{-1}\Big(- G(x_n) + G(\hat{x}) + G'(x_n)\,\hat{e}_n + (B_n - G'(x_n))\,\hat{e}_n\Big)\right\| \\
&\leq \left\|B_n^{-1}\right\| \left\| \int_0^1 \Big(G'(x_n) - G'(\hat{x} + t\,\hat{e}_n)\Big)\,\hat{e}_n\,dt + (B_n - G'(x_n)\,)\,\hat{e}_n\right\| \\
&\leq \|B_n^{-1}\|\,\|\hat{e}_n\| \left(\int_0^1 \|G'(x_n) - G'(\hat{x} + t\,\hat{e}_n)\,\|dt + \|B_n - G'(x_n)\,\|\right) \\
&\leq \|B_n^{-1}\|\,\|\hat{e}_n\| \left(\int_0^1 \frac{\eta\|x_n - \hat{x} - t\,\hat{e}_n\,\|}{r}dt + \|B_n - G'(x_n)\,\|\right) \\
&\leq \|B_n^{-1}\|\,\|\hat{e}_n\| \Big(\frac{\eta\|\hat{e}_n\|}{2r} + \|B_n - G'(x_n)\,\|\Big).
\end{aligned}$$

Since $\hat{x}\,,\ x_n \in \overline{B_r(x_0)}$ then $\|\hat{e}_n\| \leq 2r$, then by using above inequality, we have

$$\|\hat{e}_{n+1}\| \leq \beta\,\|\hat{e}_n\|\,\Big(\frac{\eta\|\hat{e}_n\|}{2r} + \alpha r\Big) \leq \beta(\eta + \alpha)\|\hat{e}_n\| \leq \frac{1}{2}\|\hat{e}_n\|,$$

if we assume $\beta(\eta + \alpha) \leq \dfrac{1}{2}$. So we have $\|x_n - \hat{x}\| \leq \dfrac{1}{2^n}$. Therefore,

$$\|\hat{x} - x^*\| \leq \|\hat{x} - x_n\| + \|x_n - x^*\| \leq \frac{1}{2^n} + r^{n+1}.$$

Let $n \to \infty$ to obtain the uniqueness result. Now assume $r \leq \dfrac{1}{2}$ and $\eta = \dfrac{1}{6(2 + r)}$, the

constants in the proof of this claims could be chosen as:

$$d = \frac{1+3r}{6(2+r)^2}, \quad \xi = \frac{11+7r}{6(2+r)^2}, \quad \beta = \frac{3(2+r)^2}{11+7r},\ \mu = \frac{12+6r}{12+11r}, \quad \alpha = \frac{3+2r}{2(2+r)^2}. \tag{2.8}$$

The calculations for finding the constants are given in Appendix A. This completes proof of the theorem. □

We now consider an example which illustrates that the constants in the above theorem cannot be arbitrary. Consider $N = 1$ with $\Omega = (0.1, 1)$ and $F(x) = x$. Clearly this trivial example has no solution in Ω. Take, for instance, $x_0 = 0.2$. Then for any $r < 0.1$, $\overline{B_r(x_0)} \subset \Omega$, Assumption (2.2) of the theorem reads $\|x_0\| \leq \xi r^2$, which cannot be satisfied for $\xi = (11+7r)(2+r)^{-2}/6$. It is not claimed that this value of ξ is optimal, but it must be sufficiently small.

Next, consider another 1D example with $\Omega = (-1, 1)$ and $F(x) = x(x+2)$. The only trouble occurs at $x = -1$ because $F'(-1) = 0$. We check the hypotheses of the above theorem for this simple example. Consider $r = 0.05$. Using (2.8), the inequality (2.2) is equivalent to $x_0 \in [-0.0011, 0.0012]$, while the inequality (2.3) becomes $x_0 \geq -0.3850$, which is less stringent than (2.2). Finally, (2.4) is equivalent to $1.5439(x_0 + 1) \leq A_0 \leq 2.0046(x_0 + 1)$. Note that the lower bound is positive and, in conjunction with (2.2), guarantees convergence of the iteration to the root 0.

Next we show superlinear convergence of Broyden's method. The proof follows closely that of Theorem 8.2.2 in [24].

Theorem 2.13. Assume the hypotheses of Theorem 2.12. Then Broyden's method converges superlinearly to a unique zero of F in $\overline{B_r(x_0)}$.

Proof. By Theorem 2.12, the iterates x_n defined by

$$\begin{aligned} &B_0 = F'(x_0)^{-1}A_0, \quad x_0, A_0 \text{ given, invertible,} \\ &x_{n+1} = x_n + s_n, \qquad s_n = -B_n^{-1}G(x_n), \qquad n \geq 0, \\ &B_{n+1} = B_n + \frac{G(x_{n+1})s_n^T}{\|s_n\|^2}\ , \end{aligned}$$

converge to x, a zero of F in $\overline{B_r(x_0)}$, where $G(z) = F'(x_0)^{-1}F(z)$. Assume that $F(x_n) \neq 0$ for all $n \geq 0$. Therefore we have:

$$\begin{aligned} &\|e_n\| \leq r^{n+1}, && \|E_n\| \leq \alpha r, && \|B_n^{-1}\| \leq \beta, \\ &\|G(x_n)\| \leq \xi r^{n+2}, && \|G'(x_n)^{-1}\| \leq \mu, \end{aligned}$$

where $e_n = x_n - x^*$, $E_n = B_n - G'(x_n)$, and the positive constants are given by (2.8).

Claim: $\forall n \geq 0$,

1. $\|e_{n+1}\| \leq \dfrac{\|e_n\|}{2}$;
2. $\|E_{n+1}\| \leq \|E_n\| + \dfrac{3\eta}{r}\|e_n\|$.

Claim 1. First notice that:

$$\begin{aligned} e_{n+1} &= x_{n+1} - x^* = x_n + s_n - x^* = e_n - B_n^{-1}G(x_n) = B_n^{-1}B_n e_n + B_n^{-1}(-G(x_n) + G(x^*)) \\ &= B_n^{-1}\Big(B_n - G'(x_n)\ \Big)\ e_n + B_n^{-1}\ \Big(\ - G(x_n) + G(x^*) + G'(x_n)e_n\Big). \end{aligned}$$

Then we have:

$$\begin{aligned}
\|e_{n+1}\| &= \left\| B_n^{-1}\Big(-G(x_n) + G(x^*) + G'(x_n)\, e_n + (B_n - G'(x_n))\, e_n \Big) \right\| \\
&\leq \left\|B_n^{-1}\right\| \left\| \int_0^1 \Big(G'(x_n) - G'(x^* + te_n) \Big)\, e_n\, dt + (B_n - G'(x_n)\,)\, e_n \right\| \\
&\leq \|B_n^{-1}\|\, \|e_n\| \left(\int_0^1 \|G'(x_n) - G'(x^* + te_n)\, \|dt + \|B_n - G'(x_n)\, \| \right) \\
&\leq \|B_n^{-1}\|\, \|e_n\| \left(\int_0^1 \frac{\eta}{r}\, \|x_n - x^* - te_n\, \|dt + \|B_n - G'(x_n)\, \| \right) \\
&\leq \|B_n^{-1}\|\, \|e_n\| \left(\int_0^1 \frac{\eta\, \|e_n\|}{r}(1-t)\, dt + \|B_n - G'(x_n)\, \| \right) \\
&\leq \beta\, \|e_n\| \left(\frac{\eta\|e_n\|}{2r} + \alpha r\right) \leq \beta\, \|e_n\| \left(\frac{\eta r^n}{2} + \alpha r\right) \\
&\leq \beta(\eta + \alpha)\|e_n\| \leq \frac{\|e_n\|}{2}.
\end{aligned}$$

Notice that we are using the fact that for any n we have $\|e_n\| \leq r^{n+1}$ and also by definition, $\beta(\eta + \alpha) = \beta\xi = \frac{1}{2}$.

Claim 2. For any $n \geq 0$ we have:

$$\begin{aligned}
E_{n+1} &= B_{n+1} - G'(x_{n+1}) = B_n + \frac{G(x_{n+1})s_n^T}{\|s_n\|^2} - G'(x_{n+1}) \\
&= B_n - G'(x_n) + \frac{G(x_{n+1})s_n^T}{\|s_n\|^2} + G'(x_n) - G'(x_{n+1}) \\
&= E_n(I - \frac{s_n s_n^T}{\|s_n\|^2}) + \frac{(G(x_{n+1}) - G(x_n)\,)s_n^T}{\|s_n\|^2} + \frac{E_n s_n s_n^T}{\|s_n\|^2} + \frac{G(x_n)s_n^T}{\|s_n\|^2} + G'(x_n) - G'(x_{n+1}),
\end{aligned}$$

but $G(x_n) = -B_n s_n$, so we have:

$$E_{n+1} = E_n(I - \frac{s_n s_n^T}{\|s_n\|^2}) + \frac{(G(x_{n+1}) - G(x_n) - G'(x_n)s_n\,)s_n^T}{\|s_n\|^2} + G'(x_n) - G'(x_{n+1}). \quad (2.9)$$

Consider the second term of the right-hand side of this equality:

$$\begin{aligned}
\| G(x_{n+1}) - G(x_n) - G'(x_n)s_n \| &\leq \int_0^1 \| \Big(G'(t\ x_{n+1} + (1-t)x_n) - G'(x_n) \Big)s_n \| dt \\
&\leq \int_0^1 \frac{\eta \| t\ x_{n+1} + (1-t)x_n - x_n \| \|s_n\|\ dt}{r} \\
&\leq \frac{\eta \|s_n\|}{r} \|x_{n+1} - x_n\| \int_0^1 t\ dt \\
&\leq \frac{\eta \|s_n\|}{2r} \Big(\|x_{n+1} - x^*\| + \|x_n - x^*\| \Big) \\
&\leq \frac{\eta \|s_n\|}{2r} \Big(\|e_{n+1}\| + \|e_n\| \Big).
\end{aligned}$$

Therefore,

$$\| G(x_{n+1}) - G(x_n) - G'(x_n)s_n \| \leq \frac{\eta \|s_n\|}{2r} \Big(\|e_{n+1}\| + \|e_n\| \Big) \leq \frac{\eta \|s_n\|}{r} \|e_n\|, \qquad (2.10)$$

and

$$\begin{aligned}
\|E_{n+1}\| &\leq \|E_n\| + \frac{\| G(x_{n+1}) - G(x_n) - G'(x_n)s_n \| \|s_n^T\|}{\|s_n\|^2} + \|G'(x_n) - G'(x_{n+1})\| \\
&\leq \|E_n\| + \frac{\| G(x_{n+1}) - G(x_n) - G'(x_n)s_n \|}{\|s_n\|} + \eta \frac{\|x_{n+1} - x_n\|}{r} \\
&\leq \|E_n\| + \frac{\eta}{r} \|e_n\| + \frac{\eta}{r} \Big(\|e_{n+1}\| + \|e_n\| \Big) \\
&= \|E_n\| + \frac{3\eta}{r} \|e_n\|.
\end{aligned}$$

Claim:

3. There is some positive integer m so that for all $n \geq m$, $\|G(x_n)\| \geq \dfrac{\|e_n\|}{4\|G'(x_n)^{-1}\|}$;
4. $\left\| E_n \Big(I - \dfrac{s_n s_n^T}{\|s_n\|^2} \Big) \right\|_F \leq \|E_n\|_F - \dfrac{1}{2\|E_n\|_F} \dfrac{\|E_n s_n\|^2}{\|s_n\|^2}$ for all $n \geq 0$;
5. $\dfrac{\|E_n s_n\|}{\|s_n\|} \to 0$ as $n \to \infty$.

Recall from (2.5),

$$\|G(x_{n+1})\| \leq \frac{\eta\|s_n\|^2}{2r} + \alpha r\|s_n\|,$$

and also

$$G(x_{n+1}) - G(x_n) = G'(x_n)s_n + \int_0^1 \Big(G'(t\ x_{n+1} + (1-t)x_n) - G'(x_n) \Big) s_n dt.$$

Then,

$$\begin{aligned}
\|G(x_n)\| &\geq \|G'(x_n)s_n\| - \|G(x_{n+1})\| - \frac{\eta}{r}\int_0^1 t\|s_n\|^2 dt \\
&\geq \frac{\|s_n\|}{\|G'(x_n)^{-1}\|} - \|G(x_{n+1})\| - \frac{\eta}{2r}\|s_n\|^2 \\
&\geq \frac{\|s_n\|}{\|G'(x_n)^{-1}\|} - \frac{\eta\|s_n\|^2}{r} - \alpha r\|s_n\| \\
&\geq \|s_n\| \Big(\frac{1}{\|G'(x_n)^{-1}\|} - \frac{\eta\|s_n\|}{r} - \alpha r \Big).
\end{aligned}$$

Notice that $\|e_{n+1}\| \leq \|e_n\|/2$, and so

$$\frac{\|e_n\|}{2} \leq \|e_n\| - \|e_{n+1}\| \leq \|s_n\| \leq \|e_{n+1}\| + \|e_n\| \leq 2\|e_n\|,$$

leading to,

$$\|G(x_n)\| \geq \frac{\|e_n\|}{2} \Big(\frac{1}{\|G'(x_n)^{-1}\|} - \frac{2\eta\|e_n\|}{r} - \alpha r \Big). \tag{2.11}$$

Since $\|e_n\| \to 0$, there is some m so that for all $n \geq m$,

$$\|e_n\| \leq \frac{r}{4\eta\|G'(x_n)^{-1}\|} - \frac{\alpha r^2}{2\eta}.$$

Then we get

$$\frac{2\eta\|e_n\|}{r} + \alpha r \leq \frac{1}{2\|G'(x_n)^{-1}\|},$$

by substituting this in the equation (2.11) we have,

$$\|G(x_n)\| \;\; \geq \;\; \frac{\|e_n\|}{4\|G'(x_n)^{-1}\|}.$$

Claim 4. For any matrix E and vectors u and v, it can be easily proved that:

$$\|E + uv^T\|_F^2 = \|E\|_F^2 + 2v^T E^T u + \|u\|^2\|v\|^2.$$

The above identity with $u = -E_n s_n$ and $v = s_n/\|s_n\|^2$ for any $n \geq 0$ gives

$$\left\| E_n - \frac{E_n s_n s_n^T}{\|s_n\|^2} \right\|_F^2 = \|E_n\|_F^2 - \frac{\|E_n s_n\|^2}{\|s_n\|^2},$$

consequently

$$\left\| E_n\left(I - \frac{s_n s_n^T}{\|s_n\|^2}\right) \right\|_F = \left(\|E_n\|_F^2 - \frac{\|E_n s_n\|^2}{\|s_n\|^2} \right)^{1/2} \leq \|E_n\|_F - \frac{1}{2\|E_n\|_F} \frac{\|E_n s_n\|^2}{\|s_n\|^2},$$

using the inequality $(a^2 + b^2)^{1/2} \leq a - b^2/(2a)$ for any $a \geq b > 0$.

Claim 5. By using (2.9), (2.10) and (2.11),

$$\begin{aligned} \|E_{n+1}\|_F &= \left\| E_n\left(I - \frac{s_n s_n^T}{\|s_n\|^2}\right) \right\|_F + \left\| \frac{(G(x_{n+1}) - G(x_n) - G'(x_n)s_n\ )s_n^T}{\|s_n\|^2} \right\|_F \\ &\quad + \|G'(x_n) - G'(x_{n+1})\|_F \\ &\leq \|E_n\|_F - \frac{1}{2\|E_n\|_F} \frac{\|E_n s_n\|^2}{\|s_n\|^2} + \frac{3\eta}{r}\sqrt{N}\|e_n\|, \end{aligned}$$

or

$$\begin{aligned}\frac{\|E_n s_n\|^2}{\|s_n\|^2} &\leq 2\|E_n\|_F(\ \|E_n\|_F - \|E_{n+1}\|_F + \frac{3\eta}{r}\ \sqrt{N}\|e_n\|)\\ &\leq 2\sqrt{N}\alpha r(\ \|E_n\|_F - \|E_{n+1}\|_F + \frac{3\eta}{r}\ \sqrt{N}\|e_n\|).\end{aligned}$$

Summing over n from 0 to m for any m, we obtain

$$\begin{aligned}\sum_{n=0}^{m}\frac{\|E_n s_n\|^2}{\|s_n\|^2} &\leq 2\sqrt{N}\alpha r(\ \|E_0\|_F - \|E_{m+1}\|_F + \frac{3\eta}{r}\ \sqrt{N}\|e_0\|\ \sum_{n=0}^{m}\frac{1}{2^n}\ )\\ &\leq 2\sqrt{N}\alpha r(\ \|E_0\|_F + \frac{3\eta}{r}\ \sqrt{N}\|e_0\|\ ).\end{aligned}$$

Since $\|E_0\|_F \leq \sqrt{N}dr$ and $\|e_0\| \leq r$,

$$\sum_{n=0}^{m}\frac{\|E_n s_n\|^2}{\|s_n\|^2} \leq 2N(dr+3\eta)\alpha r.$$

Take $m \to \infty$ to conclude that

$$\lim_{n\to\infty}\frac{\|E_n s_n\|^2}{\|s_n\|^2} = 0.$$

This completes the proof of Claim 5. Now from the Broyden's iteration, for any $n \geq 0$,

$$0 = B_n s_n + G(x_n) = E_n s_n + G'(x_n)s_n + G(x_n).$$

Therefore,

$$-G(x_{n+1}) = E_n s_n + G'(x_n)s_n - G(x_{n+1}) + G(x_n),$$

leading to

$$\frac{\|G(x_{n+1})\|}{\|s_n\|} \leq \frac{\|E_n s_n\|}{\|s_n\|} + \frac{\|G'(x_n)s_n - G(x_{n+1}) + G(x_n)\|}{\|s_n\|} \leq \frac{\|E_n s_n\|}{\|s_n\|} + \frac{\eta}{r}\|e_n\|\ ,$$

by (2.10). By using Claim 5, we have

$$\lim_{n\to\infty} \frac{\|G(x_{n+1})\|}{\|s_n\|} \quad \leq \quad \lim_{n\to\infty} \frac{\|E_n s_n\|}{\|s_n\|} + \frac{\eta}{r} \lim_{n\to\infty} \|e_n\| = 0.$$

By Claim 3, for n big enough,

$$\frac{\|G(x_{n+1})\|}{\|s_n\|} \geq \frac{1}{4\|G'(x_{n+1})^{-1}\|} \frac{\|e_{n+1}\|}{\|s_n\|} \geq \frac{1}{4\mu} \frac{\|e_{n+1}\|}{\|s_n\|} \geq \frac{1}{4\mu} \frac{\|e_{n+1}\|}{\|e_n\| + \|e_{n+1}\|}.$$

Let $c_n = \|e_{n+1}\|/\|e_n\|$. Therefore,

$$0 \quad = \quad \lim_{n\to\infty} \frac{\|G(x_{n+1})\|}{\|s_n\|} \geq \frac{1}{4\mu} \lim_{n\to\infty} \frac{\|e_{n+1}\|}{\|e_n\| + \|e_{n+1}\|} = \frac{1}{4\mu} \lim_{n\to\infty} \frac{\|c_{n+1}\|}{1 + \|c_{n+1}\|}.$$

This implies that $\lim_{n\to\infty} c_{n+1} = 0$, which is superlinear convergence. □

Appendix A

This appendix provides the mathematical calculations for finding the constants in the proof of Theorem 2.12. The relations among the constants are given by:

1. $\eta < 1$ and $dr < 1$,
2. $\xi \geq \eta + \alpha$,
3. $\alpha \geq 3\eta + d$,
4. $\mu = \dfrac{1}{1-\eta}$,
5. $\mu\alpha r < 1$,
6. $\beta = \max\{\dfrac{1}{1-dr}, \dfrac{1}{1-\eta-\alpha r}\}$,
7. $\beta\xi \leq 1$,
8. $\beta(\eta + \alpha) \leq \dfrac{1}{2}$.

By using condition 3. we have $\alpha > d$ and so

$$\beta = \max\{\frac{1}{1-dr}, \frac{1}{1-\eta-\alpha r}\} = \frac{1}{1-\eta-\alpha r}.$$

Let $\xi = \eta + \alpha$ and $\beta\xi = \dfrac{1}{2}$ so that

$$\beta = \frac{1}{2\xi} = \frac{1}{2(\eta+\alpha)}.$$

Thus

$$\frac{1}{1-\eta-\alpha r} = \frac{1}{2(\eta+\alpha)} \Rightarrow \quad \alpha = \frac{1-3\eta}{2+r}, \quad \xi = \frac{\eta r - \eta + 1}{2+r}, \quad \beta = \frac{2+r}{2(\eta r - \eta + 1)}.$$

Note that $\eta r - \eta + 1 > 0$, if $\eta < \dfrac{1}{1-r}$. Define

$$d = \alpha - 3\eta - \frac{2\eta}{2+r} = \frac{1 - 11\eta - 3\eta r}{2+r}.$$

We need to be sure that $d > 0$. It is sufficient to consider

$$0 < \eta < \min\{\frac{1}{11+3r}, \frac{1}{1-r}\} = \frac{1}{11+3r}.$$

Notice that by the expression of ξ, α, β and d we have,

$$\begin{aligned}
&\mu\alpha r < 1 \text{ if and only if } \eta < \frac{1}{1-r}, \text{ which is true;} \\
&1 - dr = \frac{2 + 11\eta r + 3\eta r^2}{2 + r} > 0 \Rightarrow dr < 1; \\
&\beta(\eta + \alpha) = \beta\xi = \frac{1}{2}.
\end{aligned}$$

In summary, with $r \leq \frac{1}{2}$, we be could choose $\eta = \frac{1}{6(2+r)}$, therefore

$$d = \frac{1+3r}{6(2+r)^2}, \quad \xi = \frac{11+7r}{6(2+r)^2}, \quad \beta = \frac{3(2+r)^2}{11+7r}, \ \mu = \frac{12+6r}{12+11r}, \quad \alpha = \frac{3+2r}{2(2+r)^2}.$$

2.5 BFGS method

Let Ω be an open set in $\mathbb{R}^N$ and a smooth $f : \Omega \to \mathbb{R}$. The problem is to find a local minimum of f in Ω. Of course, one can simply apply Broyden's method to the nonlinear system $F(x) = \nabla f(x) = 0$. However, in general, the approximate Jacobian in Broyden's method is not symmetric, clearly not an ideal situation since the exact Jacobian is symmetric. There are many ways to obtain a quasi-Newton Broyden's method where the approximate Jacobian is symmetric. The most popular is the method of BFGS. Given $x_0 \in \Omega$ and SPD initial approximate Jacobian A_0, the iteration is:

$$
\begin{aligned}
s_n &= -A_n^{-1} F(x_n), \\
x_{n+1} &= x_n + s_n, && (2.12) \\
y_n &= F(x_{n+1}) - F(x_n), && (2.13) \\
A_{n+1} &= A_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{A_n s_n s_n^T A_n}{s_n^T A_n s_n} , && (2.14)
\end{aligned}
$$

for any $n \geq 0$. Notice that consecutive approximate Jacobians differ by a rank-two matrix. Local superlinear convergence and global convergence for BFGS with line search with classical assumptions are known. The analysis is based on early work by [57] and [22]. It assumes that the starting point is sufficiently close to the solution x^* and that the initial Hessian approximation is sufficiently close to $F'(x^*)$. We will now show convergence of the BFGS method using Kantorovich-type assumptions.

Theorem 2.14. Let Ω be an open set in $\mathbb{R}^N$, $f : \Omega \to \mathbb{R}$ and $f \in C^2(\Omega)$. Let $F(x) = \nabla f(x)$ and $F'(x) = D^2 f(x)$. Assume $x_0 \in \Omega$ and $\overline{B_r(x_0)} \subset \Omega$ for some $0 < r \leq 1/2$. Suppose there are positive constants $m \leq 1$ and M such that for any $z \in \mathbb{R}^N$ and $x \in \overline{B_r(x_0)}$,

$$
m\|z\|^2 \leq z^T D^2 f(x) z \leq M\|z\|^2.
$$

Also

$$\|F'(x_0)^{-\frac{1}{2}}F(x_0)\| \leq ar^2, \tag{2.15}$$

$$\|F'(x_0)^{-\frac{1}{2}}\Big(F'(u) - F'(v)\Big)F'(x_0)^{-\frac{1}{2}}\| \leq \frac{\eta\|u-v\|}{\sqrt{r}}, \qquad \forall u, v \in \overline{B_r(x_0)}, \tag{2.16}$$

where a and η are positive constants dependent on r (to be defined later). If r is sufficiently small (satisfies (2.35)), then the BFGS iteration $\{x_n\}$ with $A_0 = F'(x_0)$ is well defined and exactly one of the following cases holds,

(i) $F(x_n) = 0$ for some $n \geq 0$.

(ii) $\{x_n\}$ converges to a unique zero of F in $\overline{B_r(x_0)}$.

Proof. First notice that by given assumptions of the theorem, $m \leq \|F'(x)\|$ for all $x \in \overline{B_r(x_0)}$, especially $\|F'(x_0)^{-1}\| \leq \dfrac{1}{m}$.

Let $G(\xi) = F'(x_0)^{-\frac{1}{2}}F(F'(x_0)^{-\frac{1}{2}}\xi)$. Observe that $G(\xi^*) = 0$ if and only if $F(x^*) = 0$, where $x^* = F'(x_0)^{-\frac{1}{2}}\xi^*$. Since $F'(x_0)^{\frac{1}{2}}$ is invertible,

$$G(\xi^*) = F'(x_0)^{-\frac{1}{2}}F(F'(x_0)^{-\frac{1}{2}}\xi^*) = 0 \;\Leftrightarrow\; F(F'(x_0)^{-\frac{1}{2}}F'(x_0)^{\frac{1}{2}}x^*) = F(x^*) = 0.$$

Also we have:

$$G'(\xi) = F'(x_0)^{-\frac{1}{2}}F'_x(F'(x_0)^{-\frac{1}{2}}\xi)F'(x_0)^{-\frac{1}{2}}.$$

We apply BFGS method for $G(\xi)$. First notice that

$$G'(\xi_0) = F'(x_0)^{-\frac{1}{2}}F'(F'(x_0)^{-\frac{1}{2}}\xi_0)F'(x_0)^{-\frac{1}{2}} = F'(x_0)^{-\frac{1}{2}}F'(x_0)F'(x_0)^{-\frac{1}{2}} = I.$$

Define $B_0 = G'(\xi_0) = I$ and $\xi_0 = F'(x_0)^{\frac{1}{2}} x_0$. Then for $n \geq 0$,

$$
\begin{aligned}
t_n &= -B_n^{-1} G(\xi_n), \\
\xi_{n+1} &= \xi_n + t_n, \\
z_n &= G(\xi_{n+1}) - G(\xi_n), \\
B_{n+1} &= B_n + \frac{z_n z_n^T}{z_n^T t_n} - \frac{B_n t_n t_n^T B_n}{t_n^T B_n t_n} \ .
\end{aligned}
$$

Assume $F(x_n) \neq 0$ for all $n \geq 0$. We apply mathematical induction for proving $\xi_n = F'(x_0)^{\frac{1}{2}} x_n$ and $B_n = F'(x_0)^{-\frac{1}{2}} A_n F'(x_0)^{-\frac{1}{2}}$ for all n. Basic step holds trivially. Assume these statements are true for some positive integer n, then

$$
\begin{aligned}
\xi_{n+1} &= \xi_n + t_n = F'(x_0)^{\frac{1}{2}} x_n + F'(x_0)^{\frac{1}{2}} s_n \\
&= F'(x_0)^{\frac{1}{2}} (x_n + s_n) = F'(x_0)^{\frac{1}{2}} x_{n+1}.
\end{aligned}
$$

Notice that by using induction hypothesis,

$$
\begin{aligned}
t_n &= -B_n^{-1} G(\xi_n) = -\left(F'(x_0)^{-\frac{1}{2}} A_n F'(x_0)^{-\frac{1}{2}}\right)^{-1} F'(x_0)^{-\frac{1}{2}} F(F'(x_0)^{-\frac{1}{2}} \xi_n) \\
&= -F'(x_0)^{\frac{1}{2}} A_n^{-1} F(F'(x_0)^{-\frac{1}{2}} \xi_n) = -F'(x_0)^{\frac{1}{2}} A_n^{-1} F(x_n) = F'(x_0)^{\frac{1}{2}} s_n.
\end{aligned}
$$

Furthermore

$$
z_n = G(\xi_{n+1}) - G(\xi_n) = F'(x_0)^{-\frac{1}{2}} (F(x_{n+1}) - F(x_n)) = F'(x_0)^{-\frac{1}{2}} y_n,
$$

then by using definition for B_{n+1}, we obtain

$$
\begin{aligned}
B_{n+1} &= F'(x_0)^{-\frac{1}{2}} A_n F'(x_0)^{-\frac{1}{2}} + F'(x_0)^{-\frac{1}{2}} \frac{y_n y_n^T}{y_n^T s_n} F'(x_0)^{-\frac{1}{2}} - F'(x_0)^{-\frac{1}{2}} \ \frac{A_n s_n s_n^T A_n}{s_n A_n s_n} \ F'(x_0)^{-\frac{1}{2}} \\
&= F'(x_0)^{-\frac{1}{2}} \left(A_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{A_n s_n s_n^T A_n}{s_n^T A_n s_n}\right) F'(x_0)^{-\frac{1}{2}} \ = F'(x_0)^{-\frac{1}{2}} A_{n+1} F'(x_0)^{-\frac{1}{2}}.
\end{aligned}
$$

By using assumptions of the theorem, it is not difficult to show, $\|G(\xi_0)\| \leq ar^2$,

$$\|G(\xi_0)\| = \|F'(x_0)^{-\frac{1}{2}}F(F'(x_0)^{-\frac{1}{2}}\xi_0)\| = \|F'(x_0)^{-\frac{1}{2}}F(x_0)\| \leq ar^2.$$

In the following, we use assumptions of theorem to give two important properties of $G(\xi)$. Let

$$\rho = mr.$$

Claim 1. $\|G'(\omega) - G'(\tau)\| \leq \frac{\eta}{\sqrt{\rho}}\|\omega - \tau\|$ for all $\omega, \tau \in \overline{B_\rho(\xi_0)}$.
For any $\omega, \tau \in B_\rho(\xi_0)$, there are $u, v \in \overline{B_r(x_0)}$ such that $u = F'(x_0)^{-\frac{1}{2}}\omega$ and $v = F'(x_0)^{-\frac{1}{2}}\tau$. By (2.16),

$$\begin{aligned}
\|G'(\omega) - G'(\tau)\| &= \|F'(x_0)^{-\frac{1}{2}}F'(F'(x_0)^{-\frac{1}{2}}\omega)F'(x_0)^{-\frac{1}{2}} - F'(x_0)^{-\frac{1}{2}}F'(F'(x_0)^{-\frac{1}{2}}\tau)F'(x_0)^{-\frac{1}{2}}\| \\
&= \|F'(x_0)^{-\frac{1}{2}}\Big(F'(u) - F'(v)\Big)F'(x_0)^{-\frac{1}{2}}\| \\
&\leq \frac{\eta}{\sqrt{r}}\|u - v\| \leq \frac{\eta}{\sqrt{r}}\|F'(x_0)^{-\frac{1}{2}}\|\|\omega - \tau\| \\
&\leq \frac{\eta}{\sqrt{mr}}\|\omega - \tau\| \leq \frac{\eta}{\sqrt{\rho}}\|\omega - \tau\|,
\end{aligned}$$

since $\|F'(x_0)^{-\frac{1}{2}}\| \leq \frac{1}{\sqrt{m}}$. This completes the proof of Claim 1.

Claim 2. $\|G(\omega) - G(\tau) - G'(\tau)(\omega - \tau)\| \leq \frac{\eta}{2\sqrt{\rho}}\|\omega - \tau\|^2$ for all $\omega, \tau \in \overline{B_\rho(\xi_0)}$.
Using Claim 1,

$$\begin{aligned}
\|G(\omega) - G(\tau) - G'(\tau)(\omega - \tau)\| &= \left\| \int_0^1 \Big(G'(t\omega + (1-t)\tau) - G'(\tau)\Big)(\omega - \tau)dt \right\| \\
&\leq \|\omega - \tau\| \int_0^1 \|G'\Big(\ t\omega + (1-t)\tau\ \Big) - G'(\tau)\|dt \\
&\leq \|\omega - \tau\| \int_0^1 \frac{\eta}{\sqrt{\rho}}\|t\omega + (1-t)\tau - \tau\|dt = \frac{\eta}{2\sqrt{\rho}}\|\omega - \tau\|^2,
\end{aligned}$$

establishing Claim 2.

Claim: There are some positive constants ζ, μ, γ and β dependent on ρ (to be defined later), such that for $n \geq 0$,

3. $\|\xi_n - \xi_0\| \leq \rho(1-\rho^n)$;

4. $\|G(\xi_n)\| \leq \zeta\rho^{n+2}$;

5. $G'(\xi_n)$ is invertible and $\|G'(\xi_n)^{-1}\| \leq \mu$;

6. B_n is invertible and $\|G'(\xi_n)^{-1} - B_n^{-1}\| \leq \gamma\rho(1-\rho^n)$;

7. $\|B_n^{-1}\| \leq \beta$;

8. $\|t_n\| \leq \rho^{n+2}$.

Notice that if $\|\xi_n - \xi_0\| \leq \rho(1-\rho^n)$, then $\|x_n - x_0\| \leq \sqrt{m}r(1-m^n r^n) < r$ since $m \leq 1$. Thus $x_n \in B_r(x_0)$ and $\xi_n \in B_\rho(\xi_0)$. Also $\|t_n\| \leq \rho^{n+2}$ results in $\|s_n\| \leq r^{n+2}$.

Now we prove Claims 3 to 8 by using induction. The base case for Claim 3 is trivial. Since $\|G(\xi_0)\| \leq ar^2$, assume $\zeta \geq a/m^2$, then $\|G(\xi_0)\| \leq \zeta\rho^2$. (*Note that all additional assumptions on constants in this proof are summarized at the beginning of Appendix 2.*) Also $\|G'(\xi_0)^{-1}\| = \|I\| = 1$. Let $\mu \geq 1$, then $\|G'(\xi_0)^{-1}\| \leq \mu$. By assumption $B_0 = I$, so it is invertible and by choosing $\beta \geq 1$, the base cases for Claim 6 and 7 are satisfied. Also $\|t_0\| = \| - B_0^{-1}G(\xi_0)\| = \|G(\xi_0)\| \leq \zeta\rho^2 \leq \rho^2$, if we require $\zeta \leq 1$.

Next, assume all of the statements are true for some integer $n \geq 1$, we will show they hold for $n+1$.

Claim 3. Since B_n is invertible by hypothesis of induction, ξ_{n+1} exists and

$$\|\xi_{n+1} - \xi_0\| \;\leq\; \|\xi_{n+1} - \xi_n\| + \|\xi_n - \xi_0\| \leq \rho^{n+2} + \rho(1-\rho^n) = \rho(1-\rho^n(1-\rho)) \leq \rho(1-\rho^{n+1}),$$

since $\rho \leq \dfrac{1}{2}$. This completes the proof of Claim 3. Moreover $\|\xi_{n+1} - \xi_0\| < \rho$, so $\xi_{n+1} \in B_\rho(\xi_0)$.

Claim 4. We first show that there is a constant α such that

$$\|G'(\xi_n) - B_n\| \leq \alpha\rho.$$

By the induction hypothesis,

$$\begin{aligned}\|I - B_n^{-1}\| &\leq \|I - G'(\xi_n)^{-1}\| + \|G'(\xi_n)^{-1} - B_n^{-1}\| \\ &\leq \|G'(\xi_n)^{-1}\| \, \|G'(\xi_0) - G'(\xi_n)\| + \|G'(\xi_n)^{-1} - B_n^{-1}\| \\ &\leq \mu\eta\sqrt{\rho}(1-\rho^n) + \gamma\rho(1-\rho^n).\end{aligned}$$

By assuming $\hat{\gamma} = \mu\eta + \gamma\sqrt{\rho}$, we get $\|I - B_n^{-1}\| \leq \hat{\gamma}\sqrt{\rho}$. Let $\lambda_j,\ 1 \leq j \leq N$ be eigenvalues of B_n^{-1}. Therefore $|1-\lambda_j| \leq \hat{\gamma}\sqrt{\rho}$ for all $1 \leq j \leq N$. Also $\|B_n\| = \max_{1\leq j\leq N}\left|\frac{1}{\lambda_j}\right| \leq \frac{1}{1-\hat{\gamma}\sqrt{\rho}}$, assuming $\hat{\gamma}\sqrt{\rho} < 1$. Then

$$\begin{aligned}\|G'(\xi_n) - B_n\| &= \|G'(\xi_n)(G'(\xi_n)^{-1} - B_n^{-1})B_n\| \\ &\leq \|G'(\xi_n)\| \, \|G'(\xi_n)^{-1} - B_n^{-1}\| \, \|B_n\| \leq \frac{\mu\gamma}{1-\hat{\gamma}\sqrt{\rho}}\,\rho \leq \alpha\rho,\end{aligned}$$

by assuming $\dfrac{\mu\gamma}{1-\hat{\gamma}\sqrt{\rho}} \leq \alpha$.
Now we proceed to prove Claim 4 by using induction. By definition, $t_n = -B_n^{-1}G(\xi_n)$. Use Claim 2 to get

$$\begin{aligned}\|G(\xi_{n+1})\| &= \|G(\xi_{n+1}) - G(\xi_n) - G'(\xi_n)t_n + G'(\xi_n)t_n - B_n t_n\| \\ &\leq \|G(\xi_{n+1}) - G(\xi_n) - G'(\xi_n)t_n\| + \|(G'(\xi_n) - B_n)t_n\| \\ &\leq \frac{\eta\|t_n\|^2}{2\sqrt{\rho}} + \alpha\rho\|t_n\| = \|t_n\|(\frac{\eta\|t_n\|}{2\sqrt{\rho}} + \alpha\rho) \\ &\leq \rho^{n+2}\,(\frac{\eta\rho^{n+2}}{2\sqrt{\rho}} + \alpha\rho) \leq \rho^{n+3}\,(\eta\sqrt{\rho} + \alpha).\end{aligned}$$

If we assume $\eta\sqrt{\rho} + \alpha \leq \zeta$, then $\|G(\xi_{n+1})\| \leq \zeta\rho^{n+3}$, as we need for Claim 4.

Claim 5. By Claim 1,

$$\|I - G'(\xi_{n+1})\| = \|G'(\xi_0) - G'(\xi_{n+1})\| \leq \frac{\eta}{\sqrt{\rho}}\|\xi_{n+1} - \xi_0\| \leq \eta\sqrt{\rho}.$$

Assume $\eta\sqrt{\rho} < 1$, then by using Lemma 2.9 , $G'(\xi_{n+1})$ is invertible and

$$\|G'(\xi_{n+1})^{-1}\| \leq \frac{1}{1-\eta\sqrt{\rho}}.$$

Define

$$\mu = \frac{1}{1-\eta\sqrt{\rho}}, \tag{2.17}$$

then $\|G'(\xi_{n+1})^{-1}\| \leq \mu$, which is Claim 5.

Claim 6. First, there is some $\tilde{\xi}$ between ξ_n and ξ_{n+1} so that

$$t_n^T z_n = t_n^T(G(\xi_{n+1}) - G(\xi_n)) = t_n^T G'(\tilde{\xi}) t_n > 0,$$

since $D^2 f$ and hence G' is SPD in a neighbourhood of the initial point. Hence B_{n+1} is invertible and, in fact, SPD. Take any k satisfying $0 \leq k \leq n$. By using Sherman-Morrison-Woodbury formula,

$$B_{k+1}^{-1} = B_k^{-1} + \frac{t_k t_k^T}{z_k^T t_k}(1 + \frac{z_k^T B_k^{-1} z_k}{z_k^T t_k}) - \frac{B_k^{-1} z_k t_k^T + t_k z_k^T B_k^{-1}}{z_k^T t_k}.$$

Define

$$P_k = I - \frac{t_k z_k^T}{t_k^T z_k},$$

then,

$$B_{k+1}^{-1} = P_k B_k^{-1} P_k^T + \frac{t_k t_k^T}{t_k^T z_k}.$$

For brevity let $B = G'(\xi_k)$. After some calculations,

$$\begin{aligned} B^{-1} - B_{k+1}^{-1} &= B^{-1} - P_k B_k^{-1} P_k^T - \frac{t_k t_k^T}{t_k^T z_k} \\ &= P_k(B^{-1} - B_k^{-1})P_k^T - \frac{(t_k - B^{-1}z_k)t_k^T + t_k(t_k - B^{-1}z_k)^T P_k^T}{t_k^T z_k}. \end{aligned}$$

Define the following norm which depends on the iteration number $k \geq 0$:

$$\|X\|_k = \|G'(\xi_k)^{1/2}\, X\, G'(\xi_k)^{1/2}\|_F,$$

for any arbitrary matrix $X \in \mathbb{R}^{N\times N}$. Observe that

$$\|B^{-1} - B_{k+1}^{-1}\|_k \leq \|P_k(B^{-1} - B_k^{-1})P_k^T\|_k + \frac{\|(t_k - B^{-1}z_k)t_k^T\|_k}{t_k^T z_k} + \frac{\|t_k(t_k - B^{-1}z_k)^T P_k^T\|_k}{t_k^T z_k}. \tag{2.18}$$

Below we will find estimations for each term of this inequality. For the first term,

$$\begin{aligned} \|P_k(B^{-1} - B_k^{-1})P_k^T\|_k &= \|B^{\frac{1}{2}} P_k B^{-\frac{1}{2}} B^{\frac{1}{2}}(B^{-1} - B_k^{-1})B^{\frac{1}{2}} B^{-\frac{1}{2}} P_k^T B^{\frac{1}{2}}\|_F \\ &\leq \|B^{\frac{1}{2}} P_k B^{-\frac{1}{2}}\|^2 \, \|B^{-1} - B_k^{-1}\|_k \\ &= \|B^{\frac{1}{2}}(I - \frac{t_k z_k^T}{t_k^T z_k})B^{-\frac{1}{2}}\|^2 \, \|B^{-1} - B_k^{-1}\|_k \\ &= \left\| I - \frac{(B^{\frac{1}{2}}t_k)(B^{-\frac{1}{2}}z_k)^T}{(B^{\frac{1}{2}}t_k)^T(B^{-\frac{1}{2}}z_k)} \right\|^2 \|B^{-1} - B_k^{-1}\|_k \\ &= \left(\frac{\|B^{\frac{1}{2}}t_k\| \, \|B^{-\frac{1}{2}}z_k\|}{(B^{\frac{1}{2}}t_k)^T(B^{-\frac{1}{2}}z_k)} \right)^2 \|B^{-1} - B_k^{-1}\|_k. \end{aligned}$$

For the last line we used Lemma 2.8. Define

$$w = \frac{(B^{\frac{1}{2}}t_k)^T(B^{-\frac{1}{2}}z_k)}{\|B^{\frac{1}{2}}t_k\| \, \|B^{-\frac{1}{2}}z_k\|} \leq 1, \tag{2.19}$$

so we obtain

$$\|P_k(B^{-1} - B_k^{-1})P_k^T\|_k \;\leq\; \frac{1}{w^2}\|B^{-1} - B_k^{-1}\|_k. \tag{2.20}$$

Consider the second term of the inequality (2.18),

$$\begin{aligned}\frac{\|(t_k - B^{-1}z_k)t_k^T\|_k}{t_k^T z_k} &= \frac{\|B^{\frac{1}{2}}(t_k - B^{-1}z_k)t_k^T B^{\frac{1}{2}}\|_F}{t_k^T z_k} \\ &= \frac{\|B^{\frac{1}{2}}(t_k - B^{-1}z_k)\|\|B^{\frac{1}{2}}t_k\|}{t_k^T z_k} \\ &= \frac{1}{w}\frac{\|B^{\frac{1}{2}}t_k - B^{-\frac{1}{2}}z_k\|}{\|B^{-\frac{1}{2}}z_k\|}.\end{aligned} \tag{2.21}$$

Similarly for the last term of (2.18),

$$\begin{aligned}\frac{\|t_k(t_k - B^{-1}z_k)^T P_k^T\|_k}{t_k^T z_k} &= \frac{\|B^{\frac{1}{2}}t_k(t_k - B^{-1}z_k)^T P_k^T B^{\frac{1}{2}}\|_F}{t_k^T z_k} \\ &= \frac{\|B^{\frac{1}{2}}t_k\Big(B^{\frac{1}{2}}t_k - B^{-\frac{1}{2}}z_k\Big)^T B^{-\frac{1}{2}}P_k^T B^{\frac{1}{2}}\|_F}{t_k^T z_k} \\ &\leq \frac{\|B^{\frac{1}{2}}t_k\|\|B^{\frac{1}{2}}t_k - B^{-\frac{1}{2}}z_k\|\|B^{-\frac{1}{2}}P_k^T B^{\frac{1}{2}}\|}{t_k^T z_k} \\ &= \frac{1}{w^2}\frac{\|B^{\frac{1}{2}}t_k - B^{-\frac{1}{2}}z_k\|}{\|B^{-\frac{1}{2}}z_k\|}.\end{aligned} \tag{2.22}$$

For finding an estimation for the right-hand side of this inequality, notice that $B = G'(\xi_k)$,

$$\begin{aligned}t_k - B^{-1}z_k &= t_k - B^{-1}\Big(G(\xi_{k+1}) - G(\xi_k)\Big) \\ &= t_k - G'(\xi_n)^{-1}\Big(G(\xi_{k+1}) - G(\xi_k) - G'(\xi_k)t_k\Big) - t_k \\ &= -B^{-1}\int_0^1 \Big(G'(\xi_k + \tau t_k) - G'(\xi_k)\Big)t_k d\tau.\end{aligned}$$

Therefore by Claim 1,

$$\|B^{\frac{1}{2}}t_k - B^{-\frac{1}{2}}z_k\| \;\leq\; \frac{\eta}{2\sqrt{\rho}}\|B^{-\frac{1}{2}}\|\|t_k\|^2. \tag{2.23}$$

Since $z_k = G(\xi_{k+1}) - G(\xi_k) = G'(\tilde{\xi})t_k$ for some $\tilde{\xi}$ between ξ_{k+1} and ξ_k, it follows that $t_k = G'(\tilde{\xi})^{-1}z_k$ and

$$\begin{aligned}
&\|t_k\| = \|G'(\tilde{\xi})^{-1}\|\|z_k\| \leq \frac{M}{m}\|z_k\| \Rightarrow \frac{1}{\|z_k\|} \leq \frac{M}{m\|t_k\|},\\
&\|z_k\| = \|B^{\frac{1}{2}}B^{-\frac{1}{2}}z_k\| \leq \|B^{\frac{1}{2}}\|\|B^{-\frac{1}{2}}z_k\|,\\
&\frac{1}{\|B^{-\frac{1}{2}}z_k\|} \leq \frac{\|B^{\frac{1}{2}}\|}{\|z_k\|} \leq \frac{M\,\|B^{\frac{1}{2}}\|}{m\|t_k\|}.
\end{aligned} \tag{2.24}$$

(2.23) and (2.24) together imply:

$$\frac{\|B^{\frac{1}{2}}t_k - B^{-\frac{1}{2}}z_k\|}{\|B^{-\frac{1}{2}}z_k\|} \quad \leq \quad \frac{M\eta}{2m\sqrt{\rho}}\|B^{-\frac{1}{2}}\|\|B^{\frac{1}{2}}\|\|t_k\|. \tag{2.25}$$

By using the assumptions of the theorem, $\frac{m}{M} \leq \|G'(\xi)\| \leq \frac{M}{m}$ for any $\xi \in B_\rho(\xi_0)$. This implies $\|B\|\|B^{-1}\| = \|G'(\xi_k)\|\|G'(\xi_k)^{-1}\| \leq (\frac{M}{m})^2$. Now choose η such that $\frac{\eta\,M^2}{m^2} \leq \sqrt{2}\,$, and define

$$\Lambda = \frac{\eta\,M^2}{\sqrt{2}\,m^2} \leq 1. \tag{2.26}$$

Then

$$\frac{\eta M}{2m\sqrt{\rho}}\|B^{-\frac{1}{2}}\|\|B^{\frac{1}{2}}\|\|t_k\| \quad \leq \quad \frac{\eta\,M^2}{2m^2\sqrt{\rho}}\,\|t_k\| = \frac{\Lambda}{\sqrt{2\rho}}\|t_k\| \leq \frac{\rho^{3/2}}{\sqrt{2}} \leq \frac{1}{\sqrt{2}}. \tag{2.27}$$

From Lemma 2.10,

$$1 - w^2 \leq \frac{\Lambda^2\|t_k\|^2}{2\rho} \leq \frac{1}{2},$$

so $w^2 \geq \frac{1}{2}$ and

$$\frac{1}{w^2} = 1 + \frac{1-w^2}{w^2} \leq 1 + 2\frac{\rho^{3/2}}{\sqrt{2}}\frac{\Lambda}{\sqrt{2\rho}}\|t_k\| = 1 + \Lambda\rho\|t_k\|. \tag{2.28}$$

Combining all estimates (2.20), (2.21) and (2.22), followed by an application of (2.27), (2.28) and (2.25), inequality (2.18) becomes

$$\begin{aligned}\|B^{-1}-B_{k+1}^{-1}\|_k &\leq \frac{1}{w^2}\|B^{-1}-B_k^{-1}\|_k+\frac{2}{w^2}\frac{\|B^{\frac{1}{2}}t_k-B^{-\frac{1}{2}}z_k\|}{\|B^{-\frac{1}{2}}z_k\|}\\ &\leq (1+\Lambda\rho\|t_k\|)\,\|B^{-1}-B_k^{-1}\|_k+\sqrt{2}(1+\Lambda\rho\|t_k\|)\frac{\Lambda}{\sqrt{\rho}}\|t_k\|. \quad (2.29)\end{aligned}$$

Notice that for any arbitrary matrix $X\in\mathbb{R}^{N\times N}$,

$$\begin{aligned}\|X\|_{k+1} &= \|G'(\xi_{k+1})^{\frac{1}{2}}\,X\,G'(\xi_{k+1})^{\frac{1}{2}}\|_F\\ &= \|G'(\xi_{k+1})^{\frac{1}{2}}G'(\xi_k)^{-\frac{1}{2}}\;G'(\xi_k)^{\frac{1}{2}}\,X\,G'(\xi_k)^{\frac{1}{2}}\;G'(\xi_k)^{-\frac{1}{2}}G'(\xi_{k+1})^{\frac{1}{2}}\|_F\\ &\leq \|G'(\xi_{k+1})G'(\xi_k)^{-1}\|\|X\|_k.\end{aligned}$$

In last line we have used Lemma 2.6. Observe that,

$$G'(\xi_{k+1})G'(\xi_k)^{-1}=(G'(\xi_{k+1})-G'(\xi_k)+G'(\xi_k))G'(\xi_k)^{-1}=(G'(\xi_{k+1})-G'(\xi_k))G'(\xi_k)^{-1}+I,$$

therefore by Claim 1,

$$\|G'(\xi_{k+1})G'(\xi_k)^{-1}\|\leq 1+\|G'(\xi_{k+1})-G'(\xi_k)\|\,\|G'(\xi_k)^{-1}\|\leq 1+\frac{\eta\mu}{\sqrt{\rho}}\|t_k\|, \qquad (2.30)$$

so we obtain

$$\|X\|_{k+1}\leq(1+\frac{\eta\mu}{\sqrt{\rho}}\|t_k\|)\,\|X\|_k,$$

Define $\kappa=1+\frac{\eta\mu}{\sqrt{\rho}}\|t_k\|$. Therefore $\|X\|_{k+1}\leq\kappa\|X\|_k$. Notice that from (2.17)

$$\kappa=1+\frac{\eta\mu}{\sqrt{\rho}}\|t_k\|\leq 1+\eta\mu\rho^{k+3/2}\leq 1+\frac{\eta\sqrt{\rho}}{1-\eta\sqrt{\rho}}=\mu.$$

From the inequality (2.29),

$$\|B^{-1} - B_{k+1}^{-1}\|_{k+1} \leq \kappa(1 + \Lambda\rho\|t_k\|)\,\|B^{-1} - B_k^{-1}\|_k + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1 + \Lambda\rho\|t_k\|)\,\|t_k\|,$$

so

$$\begin{aligned}
\|B^{-1} - B_{k+1}^{-1}\|_{k+1} - \|B^{-1} - B_k^{-1}\|_k &\leq (\kappa - 1 + \kappa\Lambda\rho\|t_k\|)\,\|B^{-1} - B_k^{-1}\|_k + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1 + \Lambda)\|t_k\| \\
&\leq \left(\frac{\eta\mu}{\sqrt{\rho}}\|t_k\| + \kappa\Lambda\rho\|t_k\|\right)\|B^{-1} - B_k^{-1}\|_k + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1 + \Lambda)\|t_k\| \\
&\leq \left(\left(\frac{\eta\mu}{\sqrt{\rho}} + \kappa\Lambda\rho\right)\|B^{-1} - B_k^{-1}\|_k + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1 + \Lambda)\right)\|t_k\|.
\end{aligned}$$

Notice that $B = G'(\xi_k)$, by adding and subtracting $G'(\xi_{k+1})$,

$$\begin{aligned}
\|G'(\xi_{k+1})^{-1} - B_{k+1}^{-1}\|_{k+1} &- \|G'(\xi_k)^{-1} - B_k^{-1}\|_k \leq \|G'(\xi_{k+1})^{-1} - G'(\xi_k)^{-1}\|_{k+1} \\
&+ \left(\left(\frac{\eta\mu}{\sqrt{\rho}} + \kappa\Lambda\rho\right)\|B^{-1} - B_k^{-1}\|_k + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1 + \Lambda)\right)\|t_k\| \qquad (2.31)
\end{aligned}$$

Also by Lemma 2.6 and (2.30)

$$\begin{aligned}
\|G'(\xi_{k+1})^{-1} - G'(\xi_k)^{-1}\|_{k+1} &= \|G'(\xi_{k+1})^{-1}G'(\xi_k)G'(\xi_k)^{-1} - G'(\xi_{k+1})^{-1}G'(\xi_{k+1})G'(\xi_k)^{-1}\|_{k+1} \\
&= \|G'(\xi_{k+1})^{-1}\,\left(G'(\xi_k) - G'(\xi_{k+1})\right)\,G'(\xi_k)^{-1}\|_{k+1} \\
&= \|G'(\xi_{k+1})^{\frac{1}{2}}G'(\xi_{k+1})^{-1}\,\left(G'(\xi_k) - G'(\xi_{k+1})\right)\,G'(\xi_k)^{-1}G'(\xi_{k+1})^{\frac{1}{2}}\|_F \\
&\leq \|G'(\xi_{k+1})^{-\frac{1}{2}}\|\|G'(\xi_k)^{-1}G'(\xi_{k+1})^{\frac{1}{2}}\|\frac{\sqrt{N}\eta}{\sqrt{\rho}}\|\xi_{k+1} - \xi_k\| \\
&\leq \|G'(\xi_{k+1})^{-\frac{1}{2}}\|\|G'(\xi_k)^{-\frac{1}{2}}\|\sqrt{\|G'(\xi_k)^{-1}G'(\xi_{k+1})\|}\frac{\sqrt{N}\eta}{\sqrt{\rho}}\|\xi_{k+1} - \xi_k\| \\
&\leq \frac{\sqrt{\kappa N}\eta\mu}{\sqrt{\rho}}\|t_k\|.
\end{aligned}$$

Then by substituting this in (2.31)

$$\begin{aligned}\|G'(\xi_{k+1})^{-1} - B_{k+1}^{-1}\|_{k+1} &- \|G'(\xi_k)^{-1} - B_k^{-1}\|_k \\ &\leq \left(\frac{\sqrt{\kappa N}\eta\mu}{\sqrt{\rho}} + (\frac{\eta\mu}{\sqrt{\rho}} + \kappa\Lambda\rho)\,\|B^{-1} - B_k^{-1}\|_k + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1+\Lambda)\right)\|t_k\|.\end{aligned}$$

From the induction hypothesis, $\|B^{-1} - B_k^{-1}\|_k \leq \gamma\rho(1-\rho^k) \leq \gamma\rho$. Take the sum from $k = 0$ to $k = n$ and using Claim 8 to obtain

$$\begin{aligned}\|G'(\xi_{n+1})^{-1} - B_{n+1}^{-1}\|_{n+1} &- \|G'(\xi_0)^{-1} - B_0^{-1}\|_0 \\ &\leq \left(\frac{\sqrt{\kappa N}\eta\mu}{\sqrt{\rho}} + (\frac{\eta\mu}{\sqrt{\rho}} + \kappa\Lambda\rho)\gamma\rho + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1+\Lambda)\right)\rho^2 \sum_{k=0}^{n}(2\rho)^k \qquad (2.32)\end{aligned}$$

Notice that $G'(\xi_0) = B_0 = I$, so $\|G'(\xi_0)^{-1} - B_0^{-1}\|_0 = 0$.

$$\begin{aligned}\|G'(\xi_{n+1})^{-1} - B_{n+1}^{-1}\| &\leq \|G'(\xi_{n+1})^{-1} - B_{n+1}^{-1}\|_F \\ &= \|G'(\xi_{n+1})^{-\frac{1}{2}}G'(\xi_{n+1})^{\frac{1}{2}}\,(G'(\xi_{n+1})^{-1} - B_{n+1}^{-1})\,G'(\xi_{n+1})^{\frac{1}{2}}G'(\xi_{n+1})^{-\frac{1}{2}}\|_F \\ &\leq \|G'(\xi_{n+1})^{-1}\|\|G'(\xi_{n+1})^{-1} - B_{n+1}^{-1}\|_{n+1}.\end{aligned}$$

Use inequality (2.32) to obtain,

$$\begin{aligned}\|G'(\xi_{n+1})^{-1} - B_{n+1}^{-1}\| &\leq \|G'(\xi_{n+1})^{-1}\|\left[\frac{\sqrt{\kappa N}\eta\mu}{\sqrt{\rho}} + (\frac{\eta\mu}{\sqrt{\rho}} + \kappa\Lambda\rho)\gamma\rho + \sqrt{2}\kappa\frac{\Lambda}{\sqrt{\rho}}(1+\Lambda)\right]\rho^2 \sum_{k=0}^{n}\rho^k \\ &\leq \mu\left[\sqrt{\kappa N}\eta\mu + (\eta\mu + \kappa\Lambda\rho^{3/2})\gamma\rho + \sqrt{2}\kappa\Lambda(1+\Lambda)\right]\rho^{3/2} \sum_{k=0}^{n}\rho^k \\ &\leq \mu^2(\sqrt{2\mu N}\eta + (\eta + \rho^{3/2})\gamma\rho + 2\sqrt{2})\rho^{3/2} \sum_{k=0}^{n}\rho^k,\end{aligned}$$

since $\Lambda \leq 1,\ \kappa \leq \mu$ and $\eta \leq \sqrt{2}$ due to (2.26). Notice that $\rho \leq \frac{1}{2}$, then

$$\sum_{k=0}^{n}\rho^k = \frac{1-\rho^{n+1}}{1-\rho} \leq 2(1-\rho^{n+1}).$$

Therefore by assuming ρ such that $4\mu^2(\sqrt{\mu N}+\gamma\rho+\sqrt{2})\sqrt{\rho}\leq\gamma$,

$$\|G'(\xi_{n+1})^{-1}-B_{n+1}^{-1}\|\leq 4\mu^2(\sqrt{\mu N}+\gamma\rho+\sqrt{2})\rho^{3/2}(1-\rho^{n+1})\leq\gamma\rho(1-\rho^{n+1}).$$

This concludes the proof of Claim 6.

Claim 7. From Claim 6,

$$\|B_{n+1}^{-1}\|\leq\gamma\rho+\|G'(\xi_{n+1})^{-1}\|\leq\gamma\rho+\mu.$$

Define $\beta\geq\gamma\rho+\mu$, then $\|B_{n+1}^{-1}\|\leq\beta$, which is Claim 7.

Claim 8. By using definition,

$$\|t_{n+1}\|=\|-B_{n+1}^{-1}G(\xi_{n+1})\|\leq\|B_{n+1}^{-1}\|\|G(\xi_{n+1})\|\leq\beta\zeta\ \rho^{n+3},$$

assume $\beta\zeta\leq 1$, then $\|t_{n+1}\|\leq\rho^{n+3}$, establishing Claim 8.

Therefore by using mathematical induction we have the results. A consequence of Claim 8 is that $\{\xi_n\}$ is a Cauchy sequence lying in $B_\rho(\xi_0)$. Given $p,q\geq 0$,

$$\|\xi_p-\xi_{p+q}\|\leq\sum_{k=p}^{p+q-1}\|\xi_{k+1}-\xi_k\|\leq\sum_{k=p}^{p+q-1}\rho^{k+2}<\rho^2\sum_{k=p}^{\infty}\rho^k=\frac{\rho^{p+2}}{1-\rho}\leq\rho^{p+1}.$$

Therefore $\{\xi_n\}$ converges to a point $\xi^*\in\overline{B_\rho(\xi_0)}$. By using the fact that G is a continuous function and $\|G(\xi_n)\|\leq\zeta\rho^{n+2}$, it follows that $G(\xi^*)=0$, which implies $F(x^*)=0$, where $x^*=F'(x_0)^{-\frac{1}{2}}\xi^*$. By taking $q\to\infty$ and $p=n$ in the above calculation, $\|\xi_n-\xi^*\|\leq\rho^{n+1}$. Let $e_n=x_n-x^*$ and $\sigma_n=\xi_n-\xi^*$, then

$$\|\sigma_n\|\leq\rho^{n+1},\qquad\|e_n\|\leq r^{n+1}.$$

Notice that $\xi^*\in\overline{B_\rho(\xi_0)}$ and $x^*\in\overline{B_r(x_0)}$.

For proof of uniqueness, let $\hat{\xi}$ be any zero of G in $\overline{B_\rho(\xi_0)}$ corresponding to a root $\hat{x}$ of

F in $\overline{B_r(x_0)}$. Below we show that $\|\hat{\sigma}_{n+1}\| \leq \dfrac{\|\hat{\sigma}_n\|}{2}$ for $n \geq 0$, where $\hat{\sigma}_n = \xi_n - \hat{\xi}$. Notice that:

$$\begin{aligned}
\hat{\sigma}_{n+1} &= \xi_{n+1} - \hat{\xi} = \xi_n + t_n - \hat{\xi} = \xi_n - B_n^{-1}G(\xi_n) - \hat{\xi} \\
&= B_n^{-1}B_n\ \hat{\sigma}_n + B_n^{-1}\Big(- G(\xi_n) + G(\hat{\xi})\ \Big) \\
&= B_n^{-1}\Big(B_n - G'(\xi_n)\ \Big)\ \hat{\sigma}_n + B_n^{-1}\ \Big(- G(\xi_n) + G(\hat{\xi}) + G'(\xi_n)\hat{\sigma}_n\Big).
\end{aligned}$$

By Claim 2,

$$\begin{aligned}
\|\hat{\sigma}_{n+1}\| &= \Big\|B_n^{-1}\left(-G(\xi_n) + G(\hat{\xi}) + G'(\xi_n)\hat{\sigma}_n + (B_n - G'(\xi_n))\hat{\sigma}_n\right)\Big\| \\
&\leq \|B_n^{-1}\|\ \|\hat{\sigma}_n\|\ \Big(\frac{\eta\|\hat{\sigma}_n\|}{2\sqrt{\rho}} + \|B_n - G'(\xi_n)\|\Big). \qquad (2.33)
\end{aligned}$$

Since $\hat{\xi}$, $\xi_n \in \overline{B_\rho(\xi_0)}$ then $\|\hat{\sigma}_n\| \leq 2\rho$, then by using above inequality, we have

$$\|\hat{\sigma}_{n+1}\| \leq \beta\ \|\hat{\sigma}_n\|\ \Big(\frac{\eta\|\hat{\sigma}_n\|}{2\sqrt{\rho}} + \alpha\rho\Big) \leq \beta\ (\eta\sqrt{\rho} + \alpha\rho)\ \|\hat{\sigma}_n\| \leq \beta(\eta + \alpha\sqrt{\rho})\|\ \hat{\sigma}_n\| \leq \frac{1}{2}\|\ \hat{\sigma}_n\|,$$

if we assume $\beta(\eta + \alpha\sqrt{\rho}) \leq \dfrac{1}{2}$. Therefore $\|\xi_n - \hat{\xi}\| \leq \dfrac{1}{2^n}$ and

$$\|\hat{\xi} - \xi^*\| \leq \|\hat{\xi} - \xi_n\| + \|\xi_n - \xi^*\| \leq \frac{1}{2^n} + \rho^{n+1}.$$

Let $n \to \infty$ to obtain the uniqueness result. Let

$$\sqrt{\rho} < \frac{\eta}{96\sqrt{2}(\sqrt{N}+1)}, \qquad \eta < \min\left\{\frac{\sqrt{2}\,m^2}{M^2}, \frac{1}{6}\right\}. \qquad (2.34)$$

with the former inequality equivalent to

$$r < \frac{\eta^2}{18432(\sqrt{N}+1)^2 m}. \qquad (2.35)$$

Define the constants as

$$\mu = \frac{1}{1-\eta\sqrt{\rho}}, \tag{2.36}$$

$$\gamma = 24\sqrt{2}(\sqrt{N}+1)\sqrt{\rho}, \tag{2.37}$$

$$\beta = \gamma\rho + \mu = 24\sqrt{2}(\sqrt{N}+1)\rho\sqrt{\rho} + \mu, \tag{2.38}$$

$$\zeta = \frac{1}{2\beta} = \frac{1}{2(\gamma\rho+\mu)}, \tag{2.39}$$

$$\alpha = 4\gamma = 96\sqrt{2}(\sqrt{N}+1)\sqrt{\rho}, \tag{2.40}$$

$$a = \frac{m^2}{2(\gamma\rho+\mu)}. \tag{2.41}$$

The calculations for finding constants are given in Appendix B. This completes proof of the theorem. □

In classical proofs of global convergence of this method, the minimum is assumed to exist and a fixed, so-called, BFGS norm can be used to estimate the difference between the approximate and exact Jacobians for all iterates. In our setting, in the proof of Claim 6, the minimum is not known apriori, hence requiring a norm which changes with each iteration. We believe that this technique is applicable in more general contexts.

While the above theorem assumes uniform convexity of f, it does not follow that the iterates converge to a minimum x^* where $F(x^*) = 0$. Consider the simple example $N = 1,\ \Omega = (0.1, 1)$ and $f(x) = x^2/2$. Clearly f has no critical point in Ω. Take, for instance, $x_0 = 0.2$. Then for any $r < 0.1$, (2.16) becomes $x_0 < ar^2$, which cannot be satisified for a defined by (2.41). Again, we do not claim that this value of a is optimal, but it must be taken sufficiently small to guarantee existence of minimum x^* which is a critical point of f. Continuing with the same example, except define $\Omega = (-1, 1)$, then $A_0 = 1,\ m = M = 1,\ \rho = r$, and we can take $\eta = 0.1$, leading to

$$\mu = \frac{1}{1-0.1\sqrt{r}}, \qquad \gamma = 48\sqrt{2}\sqrt{r}, \qquad a = \frac{1}{2\Big((1-0.1\sqrt{r})^{-1} + 48\sqrt{2}r^{3/2}\Big)}.$$

Inequality (2.15) becomes $\|x_0\| \leq ar^2$, while (2.16) holds trivially. The theorem correctly states that the BFGS iterates converge to the unique minimum superlinearly.

We remark that in assumption (2.16), we assumed $r^{-1/2}$ dependence on the right-hand side. Initially, we assumed r^{-1} dependence as in the theorem for Broyden's method, but were unable assign values to constants (similar to (2.36) to (2.41)) so that all required inequalities are satisfied.

Following closely [52], we prove superlinear convergence of the BFGS method by Kantorovich-type assumptions.

Theorem 2.15. Assume the hypotheses of Theorem 2.14. Then BFGS method converges superlinearly to a unique zero of F in $\overline{B_r(x_0)}$.

Proof. By Theorem (2.14), the iterates ξ_n defined by

$$\begin{aligned} t_n &= -B_n^{-1}G(\xi_n), \\ \xi_{n+1} &= \xi_n + t_n, \\ z_n &= G(\xi_{n+1}) - G(\xi_n), \\ B_{n+1} &= B_n + \frac{z_n z_n^T}{z_n^T t_n} - \frac{B_n t_n t_n^T B_n}{t_n^T B_n t_n} \ . \end{aligned}$$

converge to ξ^*, unique zero of F in $\overline{B_\rho(\xi_0)}$, where $\rho = mr$ and $G(\xi) = F'(x_0)^{-\frac{1}{2}}F(F'(x_0)^{-\frac{1}{2}}\xi)$. Consequently x_n converges to $x^* = F'(x_0)^{-\frac{1}{2}}\xi^*$ the unique zero of $F(x)$ in $\overline{B_r(x_0)}$. Assume that $F(\xi_n) \neq 0$ for all $n \geq 0$. So $\xi_n \neq \xi^*$. Also we have:

$$\begin{aligned} &\|G'(\xi_n) - B_n\| \leq \alpha\rho, && \|B_n^{-1}\| \leq \beta, \\ &\|G'(\xi_n)^{-1} - B_n^{-1}\| \leq \gamma\rho(1-\rho^n), && \|t_n\| \leq \rho^{n+2}, \\ &\|G'(\xi_n)^{-1}\| \leq \mu, && \|\sigma_n\| \leq \rho^{n+1}, \\ &\|G(\xi_n)\| \leq \zeta\rho^{n+2}, \end{aligned}$$

where $\sigma_n = \xi_n - \xi^*$, and the positive constants $\alpha, \beta, \gamma, \mu, \zeta$ are given by equation (2.36)

to (2.41). Using exactly the same technique as in the previous theorem to show $\|\hat{\sigma}_{n+1}\| \leq \|\hat{\sigma}_n\|/2$, we could prove that $\|\sigma_{n+1}\| \leq \frac{1}{2}\|\sigma_n\|$. Let $B = G'(\xi^*)$ and

$$\tilde{t}_n = B^{\frac{1}{2}} t_n, \qquad \tilde{y}_n = B^{-\frac{1}{2}} y_n, \qquad \tilde{B}_n = B^{\frac{1}{2}} B_n B^{\frac{1}{2}}.$$

Define

$$\cos\tilde{\theta}_n = \frac{\tilde{t}_n^T \tilde{B}_n \tilde{t}_n}{\|\tilde{t}_n\|\|\tilde{A}_n \tilde{t}_n\|}, \quad \tilde{q}_n = \frac{\tilde{t}_n^T \tilde{B}_n \tilde{t}_n}{\|\tilde{t}_n\|^2} \quad \tilde{M}_n = \frac{\|\tilde{z}_n\|^2}{\tilde{z}_n^T \tilde{t}_n}, \quad \tilde{m}_n = \frac{\tilde{z}_n^T \tilde{t}_n}{\|\tilde{t}_n\|^2}.$$

Using BFGS update formula, it follows that

$$\tilde{B}_{n+1} = \tilde{B}_n + \frac{\tilde{z}_n \tilde{z}_n^T}{\tilde{z}_n^T \tilde{t}_n} - \frac{\tilde{B}_n \tilde{t}_n \tilde{t}_n^T \tilde{B}_n}{\tilde{t}_n^T \tilde{B}_n \tilde{t}_n}\ .$$

For any SPD matrix S define $\psi(S) = \text{trace}(S) - \ln\det S$. Notice that $\psi(S) > 0$ and

$$\begin{aligned} \psi(\tilde{B}_{n+1}) &= \text{Trace}(\tilde{B}_n) + \tilde{M}_n - \frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n} - \ln\det\tilde{B}_n - \ln\tilde{m}_n + \ln\tilde{q}_n \\ &= \psi(\tilde{B}_n) + (\tilde{M}_n - \ln\tilde{m}_n - 1) + (1 - \frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n} + \ln\frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n}) + \ln\cos^2\tilde{\theta}_n. \quad (2.42) \end{aligned}$$

Also,

$$z_n - Bt_n = z_n - G'(\xi^*)t_n = \int_0^1 \Big(G'(\xi_n + t_n\tau) - G'(\xi^*)\Big)t_n d\tau,$$

or

$$\tilde{z}_n - \tilde{t}_n = B^{-\frac{1}{2}} \int_0^1 \Big(G'(\xi_n + t_n\tau) - G'(\xi^*)\Big)B^{-\frac{1}{2}}\tilde{t}_n d\tau,$$

Therefore

$$\|\tilde{z}_n - \tilde{t}_n\| \leq \|B^{-\frac{1}{2}}\|^2\|\tilde{t}_n\| \int_0^1 \|G'(\xi_n + t_n\tau) - G'(\xi^*)\|d\tau \leq \frac{\eta}{\sqrt{\rho}}\|B^{-1}\|\|\tilde{t}_n\|\|\sigma_n\|.$$

By the assumption of theorem, $\frac{m}{M} \leq \|G'(\xi)\| \leq \frac{M}{m}$ for any $\xi \in B_\rho(\xi_0)$, therefore

$$\frac{\|\tilde{z}_n - \tilde{t}_n\|}{\|\tilde{t}_n\|} \leq \tilde{c}\|\sigma_n\|, \qquad \tilde{c} = \frac{m\,\eta}{M\sqrt{\rho}}. \tag{2.43}$$

Consequently

$$(1 - \tilde{c}\sigma_n)\|\tilde{t}_n\| \leq \|\tilde{z}_n\| \leq (1 + \tilde{c}\sigma_n)\|\tilde{t}_n\|. \tag{2.44}$$

Square (2.43) and use the above inequalities to get

$$(1 - \tilde{c}\sigma_n)^2\|\tilde{t}_n\|^2 - 2\tilde{z}_n^T\tilde{t}_n + \|\tilde{t}_n\|^2 \leq \|\tilde{z}_n\|^2 - 2\tilde{z}_n^T\tilde{t}_n + \|\tilde{t}_n\|^2 \leq \tilde{c}^2\sigma_n^2\|\tilde{t}_n\|^2,$$

or

$$\tilde{z}_n^T\tilde{t}_n \geq (1 - \tilde{c}\sigma_n)\|\tilde{t}_n\|^2,$$

Therefore

$$\tilde{m}_n = \frac{\tilde{z}_n^T\tilde{t}_n}{\|\tilde{t}_n\|^2} \geq (1 - \tilde{c}\sigma_n). \tag{2.45}$$

Combine this inequality and (2.44) to obtain

$$\tilde{M}_n = \frac{\|\tilde{z}_n\|^2}{\tilde{z}_n^T\tilde{t}_n} \leq \frac{(1 + \tilde{c}\sigma_n)^2}{(1 - \tilde{c}\sigma_n)}.$$

Since $\sigma_n \to 0$, there is some constant $c \geq \tilde{c}$ such that for all sufficiently large n,

$$\tilde{M}_n \leq 1 + c\sigma_n. \tag{2.46}$$

For large enough n, $\tilde{c}\sigma_n < 1/2$ and so

$$\ln(1 - \tilde{c}\sigma_n) \geq \frac{-\tilde{c}\sigma_n}{1 - \tilde{c}\sigma_n} \geq -2\tilde{c}\sigma_n,$$

apply this for (2.45) to get

$$\ln \tilde{m}_n \geq \ln(1 - \tilde{c}\sigma_n \geq -2\tilde{c}_n > -2c\sigma_n.$$

By using above inequality and (2.46) in (2.42) we have

$$0 < \psi(\tilde{B}_{n+1}) \leq \psi(\tilde{B}_n) + 3c\sigma_n + \ln\cos^2\tilde{\theta}_n + \Big[1 - \frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n} + \ln\frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n}\Big].$$

Sum over n to arrive at

$$\sum_{n=0}^{\infty}\left(\ln\frac{1}{\cos^2\tilde{\theta}_n} + \Big[1 - \frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n} + \ln\frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n}\Big]\right) \leq \psi(\tilde{B}_0) + 3c\sum_{n=0}^{\infty}\sigma_n < \infty,$$

since $\sigma_n \to 0$ by the last theorem. The term inside the square brackets is non-positive and since $\ln\cos^{-2}\tilde{\theta}_n \geq 0$, it follows that

$$\lim_{n\to\infty}\ln\frac{1}{\cos^2\tilde{\theta}_n} = 0 = \lim_{n\to\infty}\Big[1 - \frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n} + \ln\frac{\tilde{q}_n}{\cos^2\tilde{\theta}_n}\Big],$$

implying that

$$\lim_{n\to\infty}\cos^2\tilde{\theta}_n = 1 = \lim_{n\to\infty}\tilde{q}_n. \tag{2.47}$$

Therefore

$$\begin{aligned}
\lim_{n\to\infty}\frac{\|B^{-\frac{1}{2}}(B_n - B)t_n\|^2}{\|B^{\frac{1}{2}}t_n\|^2} &= \lim_{n\to\infty}\frac{\|(\tilde{B}_n - I)\tilde{t}_n\|^2}{\|\tilde{t}_n\|^2} \\
&= \lim_{n\to\infty}\frac{\|\tilde{B}_n\tilde{t}_n\|^2 - 2\tilde{t}_n^T\tilde{B}_n\tilde{t}_n + \|\tilde{t}_n\|^2}{\|\tilde{t}_n\|^2} \\
&= \lim_{n\to\infty}\frac{\tilde{q}_n^2}{\cos^2\tilde{\theta}_n} - 2\tilde{q}_n + 1 = 0.
\end{aligned}$$

Since B id SPD, it follows that

$$\lim_{n\to\infty} \frac{\|(B_n - B)t_n\|^2}{\|t_n\|^2} = 0.$$

This is sufficient to show $\xi_n \to \xi^*$ superlinearly. Let $t_n^N = -G'(\xi_n)^{-1}G(\xi_n)$ denote the Newton's step. Then

$$\begin{aligned}
\|t_n - t_n^N\| &= \|G'(\xi_n)^{-1}(G'(\xi_n)t_n + G(\xi_n))\| \\
&= \|G'(\xi_n)^{-1}\| \, \|(G'(\xi_n) - B_n)t_n\| \\
&\leq \mu\bigg(\|(G'(\xi_n) - B)t_n\| + \|(B - B_n)t_n\|\bigg) \\
&\leq \mu\bigg(\frac{\eta}{\rho}\|\sigma_n\|\|t_n\| + \|(B - B_n)t_n\|\bigg).
\end{aligned}$$

Then

$$\begin{aligned}
\lim_{n\to\infty} \frac{\|t_n - t_n^N\|}{\|t_n\|} &\leq \mu\bigg(\lim_{n\to\infty} \frac{\eta}{\rho}\|\sigma_n\| + \lim_{n\to\infty} \frac{\|(B - B_n)t_n\|}{\|t_n\|}\bigg) \\
&\leq \mu\bigg(\lim_{n\to\infty} \frac{\|(B - B_n)t_n\|}{\|t_n\|}\bigg) = 0.
\end{aligned}$$

Therefore

$$\begin{aligned}
\|\sigma_{n+1}\| &= \|\xi_n + t_n - \xi^*\| \\
&\leq \|\xi_n + t_n^N - \xi^*\| + \|t_n - t_n^N\| \\
&\leq \|\xi_n + \xi_{n+1}^N - \xi_n - \xi^*\| + \|t_n - t_n^N\| \\
&\leq \|\xi_{n+1}^N - \xi^*\| + \|t_n - t_n^N\|,
\end{aligned}$$

where ξ_{n+1}^N denotes the next iteration after ξ_n by using Newton's step. Since Newton's

method converges quadratically in a neighborhood of ξ^*, then

$$\begin{aligned}
\frac{\|\sigma_{n+1}\|}{\|\sigma_n\|} &\leq \frac{\|\xi_{n+1}^N - \xi^*\| + \|t_n - t_n^N\|}{\|\sigma_n\|} \\
&\leq \frac{\|e_{n+1}^N\|}{\|\sigma_n\|} + \frac{\|t_n - t_n^N\|\|t_n\|}{\|t_n\|\|\sigma_n\|} \\
&\leq \frac{C\|e_n^N\|^2}{\|\sigma_n\|} + \frac{\|t_n - t_n^N\|\|t_n\|}{\|t_n\|\|\sigma_n\|} \\
&\leq C\|\sigma_n\| + 2\frac{\|t_n - t_n^N\|}{\|t_n\|},
\end{aligned}$$

since $e_n^N = \xi_n - \xi^* = \sigma_n$ and $\|t_n\| \leq 2\|\sigma_n\|$. Therefore

$$\lim_{n\to\infty} \frac{\|\sigma_{n+1}\|}{\|\sigma_n\|} = C \lim_{n\to\infty} \|\sigma_n\| + 2 \lim_{n\to\infty} \frac{\|t_n - t_n^N\|}{\|t_n\|} = 0.$$

This proves superlinear convergence. □

Appendix B

This appendix provides the mathematical calculations for finding the constants in proof of Theorem 2.14. The relations among the constants could be summarized as:

1. $\rho = mr \leq \frac{1}{2}$ and $m \leq 1$,
2. $\eta < \min\left\{\frac{\sqrt{2}\,m^2}{M^2}, \frac{1}{6}\right\}$,
3. $a \leq m^2\zeta$,
4. $\mu = \frac{1}{1-\eta\sqrt{\rho}}$,
5. $4\mu^2(\sqrt{\mu N} + \gamma\rho + \sqrt{2})\sqrt{\rho} \leq \gamma$,
6. $\hat{\gamma} = \mu\eta + \gamma\sqrt{\rho}$ and $\hat{\gamma}\sqrt{\rho} < 1$,
7. $\frac{\mu\gamma}{1-\hat{\gamma}\sqrt{\rho}} \leq \alpha$,
8. $\beta \geq \max\{\gamma\rho + \mu, 1\}$,
9. $\beta\zeta \leq 1$,
10. $\eta + \alpha \leq \zeta \leq 1$,
11. $\beta(\eta + \alpha) \leq \frac{1}{2}$.

By assuming the value of the constants $\mu, \gamma, \beta, \zeta, \alpha$ and a as defined in (2.36) to (2.41), we need to show that Conditions 1 to 11 could be fulfilled if

$$\sqrt{\rho} < \frac{\eta}{96\sqrt{2}\,(\sqrt{N}+1)}, \quad \text{and} \quad \eta < \min\left\{\frac{\sqrt{2}\,m^2}{M^2}, \frac{1}{6}\right\} < \min\left\{\frac{\sqrt{2}\,m^2}{M^2}, \frac{1}{5+\sqrt{\rho}}\right\}.$$

Notice that by the last assumption,

$$\eta < \frac{1}{5+\sqrt{\rho}} < 1, \tag{2.48}$$

and also

$$\sqrt{\rho} < \frac{\eta}{96\sqrt{2}\,(\sqrt{N}+1)} < \frac{1}{8} < \frac{1-\sqrt{2/3}}{\eta}. \tag{2.49}$$

Therefore

$$(1-\eta\sqrt{\rho})^2 - 4\rho\sqrt{\rho} > (1-\eta\sqrt{\rho})^2 - \frac{1}{2} > \frac{2}{3} - \frac{1}{2} = \frac{1}{6}.$$

Use this and definition of γ given by (2.37).

$$\gamma = 24\sqrt{2}\,(\sqrt{N}+1)\sqrt{\rho} \geq 24\,(\sqrt{\mu N}+\sqrt{2})\sqrt{\rho} \geq \frac{4\,(\sqrt{\mu N}+\sqrt{2})\sqrt{\rho}}{(1-\eta\sqrt{\rho})^2-4\rho\sqrt{\rho}},$$

since by assumption (2.49), we have $\mu \leq 2$. Thus Condition 5 is satisfied. By definition of β and ζ given by (2.38) and (2.39), Conditions 8 and 9 are satisfied trivially. Observe that (2.49) implies that

$$\gamma = 24\sqrt{2}\,(\sqrt{N}+1)\sqrt{\rho} < \frac{\eta}{4}. \tag{2.50}$$

Moreover,

$$\begin{aligned} \gamma\rho < \frac{\eta\rho}{4} < \frac{\eta\sqrt{\rho}}{4} \leq \frac{\sqrt{\rho}}{20} &= \eta\sqrt{\rho}\,\frac{1}{20\,\eta} \\ &\leq \eta\sqrt{\rho}\,\Big(\frac{1}{4\eta}-\frac{1}{1-\eta\sqrt{\rho}}\Big) \qquad (2.51) \\ &\leq \frac{1}{2}-\frac{\eta\sqrt{\rho}}{1-\eta\sqrt{\rho}}. \qquad (2.52) \end{aligned}$$

In last two lines we applied (2.48) and (2.49). Since $\eta\sqrt{\rho} \leq 1$, from (2.51),

$$\gamma\rho < \frac{1}{4\eta}-\frac{1}{1-\eta\sqrt{\rho}}, \tag{2.53}$$

which results in $\zeta-\eta > \eta$ and therefore by using (2.50)

$$\alpha = 4\gamma < \eta < \zeta-\eta. \tag{2.54}$$

So Conditions 10 and 11 are satisfied. Also from (2.52)

$$\gamma\rho < \frac{1}{2}-\frac{\eta\sqrt{\rho}}{1-\eta\sqrt{\rho}}, \tag{2.55}$$

which results in $1 - \hat{\gamma}\sqrt{\rho} > \frac{1}{2}$, and

$$\frac{\mu\gamma}{1 - \hat{\gamma}\sqrt{\rho}} < 2\mu\gamma \leq 4\gamma = \alpha. \tag{2.56}$$

Thus Conditions 6 and 7 hold. Finally by using relations between constants and definition of a in (2.41), Condition 3 is satisfied.

3

Quasi-Newton methods, NCG and scaled memoryless BFGS methods

In this chapter we demonstrate local superlinear convergence of the symmetric scaled Perry NCG method and generalized scaled memoryless BFGS method by using Kantorovich-type assumptions.

3.1 Introduction

The nonlinear conjugate gradient (NCG) method is one of the most famous methods for solving unconstrained optimization problem,

$$\text{minimize} \quad f(x),$$

where $f : \mathbb{R}^N \to \mathbb{R}$ is at least continuously differentiable. It generates a sequence $\{x_n\}$ by using the iterative scheme

$$x_{n+1} = x_n + \alpha_n d_n \quad \text{for} \ \ n \geq 0, \tag{3.1}$$

where $\alpha_n > 0$ is the step length obtained by using some line search and $d_n \in \mathbb{R}^N$ is the search direction generated by

$$d_{n+1} = -g_{n+1} + \beta_n d_n, \tag{3.2}$$

where g_n denotes $F(x_n) = \nabla f(x_n)$ and β_n is a scalar parameter. The initial search direction is given by $d_0 = -g_0$. Typically, d_n is a descent direction, i.e., $g_n^T d_n < 0$. The step length α_n is chosen to give a substantial reduction of the objective function. One way is exact line search, which finds a global minimizer of the function ϕ defined by

$$\phi(\alpha) = f(x_n + \alpha d_n), \qquad \alpha > 0.$$

A more practical method applies an inexact line search to identify a step length that satisfies some adequate rules. A popular one is the Wolfe conditions,

$$f(x_n + \alpha_n d_n) \leq f(x_n) + c_1 \alpha_n g_n^T d_n,$$
$$g(x_n + \alpha_n d_n)^T d_n \geq c_2 g_n^T d_n,$$

with $0 < c_1 < c_2 < 1$. The first condition is called Armijo condition and assures sufficient decrease in the objective function and it is not enough by itself to ensure reasonable progress in a given algorithm. The second condition is called curvature condition. The strong Wolfe conditions require α_n to satisfy:

$$f(x_n + \alpha_n d_n) \leq f(x_n) + c_1 \alpha_n g_n^T d_n, \tag{3.3}$$
$$|g(x_n + \alpha_n d_n)^T d_n| \leq c_2 |g_n^T d_n|, \tag{3.4}$$

with $0 < c_1 < c_2 < 1$. In the second condition, we modify the curvature condition to force α_n to lie in a broad neighborhood of a local minimizer.

Some well known formulas for NCG parameter are β_n^{FR}, introduced by Fletcher and Reeves [28], β_n^{HS}, given by Hestenes-Stiefel [66], β_n^{PR} proposed by Polak and Ribiere [52], β_n^{CD} introduced by Fletcher [27], β_n^{LS} proposed by Liu and Storey [47] and β_n^{DY} given by Dai and Yuan [19]:

$$\beta_n^{FR} = \frac{\|g_{n+1}\|^2}{\|g_n\|^2}, \qquad \beta_n^{HS} = \frac{g_{n+1}^T y_n}{d_n^T y_n}, \qquad \beta_n^{PR} = \frac{g_{n+1}^T y_n}{\|g_n\|^2},$$

$$\beta_n^{CD} = -\frac{\|g_{n+1}\|^2}{g_n^T d_n}, \qquad \beta_n^{LS} = -\frac{g_{n+1}^T y_n}{g_n^T d_n}, \qquad \beta_n^{DY} = \frac{\|g_{n+1}\|^2}{d_n^T y_n},$$

where $y_n = g_{n+1} - g_n$. For global convergence of these methods with line search see [1], [33], [58], [59], [64], [82]. In order to accelerate efficiency of NCG methods, some authors tried to redefine them as quasi-Newton method by using secant conditions and search direction of the quasi-Newton method [3], [18], [56], [75], [81]. These methods do not necessarily generate descent directions. This can be overcome by enforcing sufficient restricting secant conditions. The well developed machinery of the latter methods have been used to analysis convergence of the formulas. For a complete survey on development on sufficient descent NCG method see [51] and the references therein.

In order to obtain a more efficient algorithm for large scale problems, memoryless BFGS method and self-scaling memoryless BFGS method are introduced. The purpose of these methods is to improve the condition number of the successive approximations to the inverse Hessian matrix. For more work on this topic see [3],[4], [5], [43].

The main purpose of this chapter is to extend the idea of using Kantorovich-type assumptions for showing superlinear local convergence of the symmetric scaled Perry NCG method and also generalized scaled memoryless BFGS method. Existence of the minimizer and the superlinear convergence are deductions of the theory. Following [15], we try to construct a simple theory with as few constants as possible.

We make the following assumptions on the objective function.

Assumption 3.1. Let Ω be an open set in $\mathbb{R}^N$, $f : \Omega \to \mathbb{R}$ and $f \in C^2(\Omega)$. Let

$F(x) = \nabla f(x)$ and $F'(x) = D^2 f(x)$. Assume $x_0 \in \Omega$ and $\overline{B_r(x_0)} \subset \Omega$ for some $0 < r \leq 1/2$. Suppose there are positive constants $m \leq 1 \leq M$ such that for any $z \in \mathbb{R}^N$ and $x \in \overline{B_r(x_0)}$,

$$m\|z\|^2 \leq z^T D^2 f(x) z \leq M\|z\|^2.$$

Also, assume

$$\begin{aligned} &\|F(x_0)\| \leq \zeta r^2, \\ &\|F'(u) - F'(v)\| \leq L\|u - v\|, \qquad \forall u, v \in \overline{B_r(x_0)}, \end{aligned}$$

where ζ and L are positive constants.

It should be emphasized that existence of a minimizer of f in Ω does not follow from Assumption 3.1. For instance, take $N = 1,\ f(x) = x^2/2,\ \Omega = (0.1, 1),\ x_0 = 0.3,\ r = 0.1$ and $\zeta \geq 30$. Then Assumption 3.1 holds and clearly f has no minimum in Ω.

Lemma 3.2 (Dennis and Schnabel 1996, [24], Lemma 4.1.12)**.** Let Ω be an open convex set in $\mathbb{R}^N$, $f : \Omega \to \mathbb{R}$ and $f \in C^2(\Omega)$. Let $F(x) = \nabla f(x)$. If for all $u, v \in \Omega$, there is some positive constant L such that

$$\|F'(u) - F'(v)\| \leq L\|u - v\|,$$

then

$$\|F(u) - F(v) - F'(v)(u - v)\| \leq \frac{L}{2}\|u - v\|^2.$$

In Section 3.2, local superlinear convergence of the symmetric scaled Perry NCG method is given using Kantorovich-type assumptions. Existence of the minimizer and the superlinear convergence are deductions of the theory. This is followed by applying a

similar analysis for a generalized scaled memoryless BFGS method in Section 3.3.

3.2 Symmetric scaled Perry NCG method

Some authors consider conjugate gradient methods as special types of quasi-Newton methods, leading them to define a new conjugacy condition for general twice continuously differentiable objective functions. Perry in [56] tried to combine the second-order information of the objective functions in NCG method to accelerate it, leading to the Perry conjugacy condition. In order to minimize a strictly convex quadratic function $f(x) = \frac{1}{2}x^T Ax - b^T x$, where A is SPD matrix, the linear conjugate gradient method generates a search direction that satisfies the conjugacy condition,

$$d_i^T A d_j = 0, \qquad \forall i \neq j.$$

Assume that f is twice continuously differentiable, define the secant condition and the quasi-Newton condition by, respectively,

$$B_{n+1}s_n = y_n \quad \text{and} \quad B_{n+1}d_{n+1} = -g_{n+1},$$

where B_{n+1} is a symmetric approximation matrix to the Hessian $F'(x_{n+1})$, $y_n = g_{n+1} - g_n$ and $s_n = x_{n+1} - x_n = \alpha_n d_n$. We have:

$$d_{n+1}^T y_n = d_{n+1}^T B_{n+1} s_n = (B_{n+1}d_{n+1})^T s_n = -g_{n+1}^T s_n.$$

Now define the Perry conjugacy condition by

$$d_{n+1}^T y_n = -g_{n+1}^T s_n.$$

This idea to define the conjugacy condition for general objective functions is due to Perry [56]. Dai and Liao [18] suggested the following generalization,

$$d_{n+1}^T y_n = -\eta_n \, g_{n+1}^T s_n, \tag{3.5}$$

where η_n is a nonnegative parameter. Then by substituting this in (3.2) we have,

$$\beta_n^P = \frac{g_{n+1}^T (y_n - \eta_n \, s_n)}{d_n^T y_n}, \tag{3.6}$$

which is called the scaled Perry NCG parameter. By employing the scaled Perry NCG parameter (3.6) in (3.2) and doing simple algebraic calculations, we obtain the search direction

$$d_{n+1} = -\left[I - \frac{s_n y_n^T}{s_n^T y_n} + \eta_n \, \frac{s_n s_n^T}{s_n^T y_n} \right] g_{n+1} \equiv -\hat{Q}_{n+1} g_{n+1}.$$

Observe that $\hat{Q}_{n+1}$ is non-symmetric.

One symmetric scaled Perry NCG approximation to the inverse Hessian is

$$Q_{n+1} = I - \frac{s_n y_n^T + y_n s_n^T}{s_n^T y_n} + \eta_n \, \frac{s_n s_n^T}{s_n^T y_n}. \tag{3.7}$$

[78] showed that if $\eta_n > \dfrac{y_n^T y_n}{s_n^T y_n}$, then Q_{n+1} is SPD and so $d_{n+1} = -Q_{n+1} g_{n+1}$ is a descent direction. The same paper gave a proof of global convergence of the symmetric scaled Perry NCG method using the Wolfe line search for general functions whose level set with respect to the initial iterate is bounded.

Notice that by selecting

$$\eta_n = 1 + \frac{y_n^T y_n}{y_n^T s_n}, \tag{3.8}$$

in (3.7), we recover the memoryless BFGS method [52].

Below we give a proof of local convergence of the symmetric scaled Perry NCG iteration using Kantorovich-type assumptions. We first offer a basic convergence theory for this class of NCG methods using the above assumptions where $F'(x_0) = I$. This assumption assures that there is a positive constant r_0 sufficiently small such that for $r \leq r_0$ and all $x \in \overline{B_r(x_0)}$,

$$m\|z\|^2 \leq z^T D^2 f(x) z \leq M\|z\|^2,$$

for any $z \in \mathbb{R}^N$ and in which m, M are positive numbers. This result comes from the continuity of eigenvalues under small perturbations [42]. However we assume the above inequalities explicitly in order to find explicit values for the constants which appear in the following theorems. We also have a parallel theory in which the assumption $F'(x_0) = I$ is generalized to $F'(x_0)$ is SPD. To obtain a convergence theory, we make a simple change of variable leading to modified NCG algorithm. Also we take the simplest version without line search for a range of η_n leading to a positive definite Q_{n+1}. With initial guess x_0 and $Q_0 = I$, the iteration is given by

$$x_{n+1} = x_n + s_n, \qquad s_n = -Q_n F(x_n), \qquad n \geq 0, \tag{3.9}$$

with the next approximation of inverse Hessian, Q_{n+1}, given by (3.7).

Theorem 3.3. Suppose that Assumption 3.1 holds with ζ given by (3.24) and L satisfies (3.23). Assume $F'(x_0) = I$. If r satisfies (3.23), then the NCG iteration $\{x_n\}$ defined by (3.7) and (3.9) for all η_n satisfying (3.18), is well defined and either $F(x_n) = 0$ for some $n \geq 0$ or $\{x_n\}$ converges to a unique zero of F in $\overline{B_r(x_0)}$.

Proof. We prove following result by using induction on $n \geq 0$.

Claim 1. There are some positive constants γ and α dependent on r (to be defined later), such that for $n \geq 0$,

(i) $\|x_n - x_0\| \leq r^{\frac{3}{2}}(1 - r^n)$;

(ii) $\|I - Q_n\| \leq \gamma r$, Q_n is invertible and $\|I - Q_n^{-1}\| \leq \alpha r$;

(iii) $\|F(x_n)\| \leq \zeta r^{n+2}$;

(iv) $\|s_n\| \leq r^{n+2}$.

The base case for Claims (i), (ii) and (iii) are trivial. Also if $\zeta \leq 1$, the base case for Claim (iv) holds.

Next, assume all of the statements are true for some integer $n \geq 1$, we will show they hold for $n + 1$.

Claim (i). Since Q_n exists, so does x_{n+1} and

$$\|x_{n+1} - x_0\| \leq \|x_{n+1} - x_n\| + \|x_n - x_0\| \leq r^{n+2} + r^{\frac{3}{2}}(1 - r^n) = r^{\frac{3}{2}}(1 - r^n(1 - \sqrt{r})) \leq r^{\frac{3}{2}}(1 - r^{n+1}).$$

In last inequality we have assumed $r \leq \frac{3-\sqrt{5}}{2}$, therefore $r \leq 1 - \sqrt{r}$. Moreover we have $\|x_{n+1} - x_0\| < r$, so $x_{n+1} \in B_r(x_0)$.

Claim (ii). Notice that there is some $\tilde{x}$ between x_n and x_{n+1} so that

$$s_n^T y_n = s_n^T(F(x_{n+1}) - F(x_n)) = s_n^T F'(\tilde{x}) s_n > 0,$$

since F' is SPD in a neighborhood of the initial point. It has recently been proved in [78] that Q_{n+1} has $N - 2$ eigenvalues equal to 1 and two other eigenvalues λ^+ and λ^- satisfy $0 < \lambda^- < \lambda^+$, if $\eta_n > \dfrac{y_n^T y_n}{s_n^T y_n}$, which results in Q_{n+1} being SPD. The symmetric scaled Perry NCG update formula (3.7) could be simplified as

$$Q_{n+1} - I = \frac{s_n((\eta_n - 1)s_n - y_n)^T + (s_n - y_n)s_n^T}{s_n^T y_n}.$$

Therefore

$$\|I - Q_{n+1}\|_F \leq \frac{\|s_n((\eta_n - 1)s_n - y_n)^T\|_F}{s_n^T y_n} + \frac{\|(s_n - y_n)s_n^T\|_F}{s_n^T y_n}. \tag{3.10}$$

Also notice that,

$$\frac{\|(s_n - y_n)s_n^T\|_F}{s_n^T y_n} = \frac{\|s_n - y_n\|\|s_n\|}{s_n^T y_n} = \frac{1}{w}\frac{\|s_n - y_n\|}{\|y_n\|},$$

where $w \equiv \dfrac{s_n^T y_n}{\|s_n\|\ \|y_n\|} \leq 1$. Moreover

$$\frac{\|((\eta_n - 1)s_n - y_n)s_n^T\|_F}{s_n^T y_n} = \frac{\|(\eta_n - 1)s_n - y_n\|\|s_n\|}{s_n^T y_n} = \frac{1}{w}\frac{\|(\eta_n - 1)s_n - y_n\|}{\|y_n\|}.$$

Therefore

$$\|I - Q_{n+1}\|_F \leq \frac{1}{w}\Big(\frac{\|s_n - y_n\|}{\|y_n\|} + \frac{\|(\eta_n - 1)s_n - y_n\|}{\|y_n\|}\Big). \tag{3.11}$$

For finding an estimation for the first term of this inequality, notice that $F'(x_0) = I$ and

$$s_n - y_n = s_n - \Big(F(x_{n+1}) - F(x_n)\Big) = -\Big(F(x_{n+1}) - F(x_n) - F'(x_0)s_n\Big).$$

Now let $\sigma_n = x_n - x_0$ and $\hat{\sigma}_n = \max\{\|\sigma_{n+1}\|, \|\sigma_n\|\}$. Then,

$$\begin{aligned}
\|s_n - y_n\| &= \|F(x_{n+1}) - F(x_n) - F'(x_0)s_n\| \\
&= \Big\| \int_0^1 \Big(F'(x_n + \tau s_n) - F'(x_0)\Big)s_n \, d\tau \Big\| \\
&\leq \|s_n\| \int_0^1 \|F'(x_n + \tau s_n) - F'(x_0)\| \, d\tau \\
&\leq L\|s_n\| \int_0^1 \|\tau(x_{n+1} - x_0) + (1 - \tau)(x_n - x_0)\| \, d\tau \\
&\leq \frac{L}{2}\|s_n\| \Big(\|\sigma_{n+1}\| + \|\sigma_n\|\Big) \leq L\hat{\sigma}_n\|s_n\|.
\end{aligned} \tag{3.12}$$

Also notice $y_n = F(x_{n+1}) - F(x_n) = F'(\tilde{x})s_n$ for some $\tilde{x}$ between x_{n+1} and x_n, it follows that $s_n = F'(\tilde{x})^{-1}y_n$ and

$$\|s_n\| \leq \|F'(\tilde{x})^{-1}\|\|y_n\| \leq \frac{\|y_n\|}{m} \quad \Rightarrow \quad \frac{1}{\|y_n\|} \leq \frac{1}{m\|s_n\|}. \tag{3.13}$$

Then (3.12) and (3.13) together imply:

$$\frac{\|s_n - y_n\|}{\|y_n\|} \leq \Lambda \hat{\sigma}_n, \tag{3.14}$$

where $\Lambda \equiv \dfrac{L}{m}$. Next to find an estimation for $\dfrac{1}{w}$, notice that by Claim 1(i), $\hat{\sigma}_n \leq r^{\frac{3}{2}}$, and

$$\|s_n - y_n\| \leq \Lambda \hat{\sigma}_n \|y_n\| \leq \Lambda r^{\frac{3}{2}} \|y_n\|.$$

Choose L such that $\Lambda = \dfrac{L}{m} \leq \dfrac{1}{2}$, then by using Lemma 2.10,

$$1 - w^2 \leq \Lambda^2 \hat{\sigma}_n^2 \leq \Lambda^2 r^3 \leq \frac{1}{2}, \tag{3.15}$$

so $w^2 \geq \dfrac{1}{2}$ and

$$\frac{1}{w} \leq \frac{1}{w^2} = 1 + \frac{1 - w^2}{w^2} \leq 1 + 2\Lambda^2 \hat{\sigma}_n^2 \leq 1 + \Lambda \hat{\sigma}_n^2. \tag{3.16}$$

For finding an estimation for the second term of the inequality (3.11), notice that

$$\begin{aligned} (\eta_n - 1)s_n - y_n &= (\eta_n - 1)s_n - (F(x_{n+1}) - F(x_n)) \\ &= -(F(x_{n+1}) - F(x_n) - F'(x_0)s_n\Big) + (\eta_n - 2)s_n. \end{aligned}$$

Then by same procedure applied in proof of (3.12),

$$\|(\eta_n - 1)s_n - y_n\| \leq \Big(L\hat{\sigma}_n + |\eta_n - 2|\Big)\|s_n\|. \tag{3.17}$$

In the following we define a restriction on η_n to find a bound for $|\eta_n - 2|$. Let η_n be such that

$$\min\{2, \frac{2y_n^T y_n}{y_n^T s_n}\} \leq \eta_n \leq \max\{2, \frac{2y_n^T y_n}{y_n^T s_n}\}. \tag{3.18}$$

We show below that η_n must satisfy the following inequality

$$|\eta_n - 2| \leq 2(1 + \Lambda\hat{\sigma}_n^2)L\hat{\sigma}_n. \tag{3.19}$$

If $\frac{y_n^T y_n}{y_n^T s_n} > 1$, then inequality (3.18) becomes

$$2 \leq \eta_n \leq \frac{2y_n^T y_n}{y_n^T s_n},$$

and

$$\begin{aligned} |\eta_n - 2| = \eta_n - 2 &\leq 2(\frac{y_n^T y_n}{y_n^T s_n} - 1) = 2\frac{y_n^T(y_n - s_n)}{y_n^T s_n} \\ &\leq 2\,\frac{\|y_n\|\,\|s_n - y_n\|}{y_n^T s_n} = \frac{2}{w}\frac{\|s_n - y_n\|}{\|s_n\|} \leq 2(1 + \Lambda\hat{\sigma}_n^2)L\hat{\sigma}_n. \end{aligned}$$

In the last line we have applied inequalities (3.12) and (3.16). If $\frac{y_n^T y_n}{y_n^T s_n} \leq 1$, then by assumption (3.18),

$$\frac{2y_n^T y_n}{y_n^T s_n} \leq \eta_n \leq 2,$$

and with similar calculations as before,

$$|\eta_n - 2| = 2 - \eta_n \leq 2(1 + \Lambda\hat{\sigma}_n^2)L\hat{\sigma}_n.$$

Applying inequality (3.19) in (3.17), we obtain

$$\|(\eta_n - 1)s_n - y_n\| \leq L\hat{\sigma}_n(3 + 2\Lambda\hat{\sigma}_n^2)\|s_n\|.$$

Then by inequality (3.13),

$$\frac{\|(\eta_n - 1)s_n - y_n\|}{\|y_n\|} \leq \Lambda\hat{\sigma}_n(3 + 2\Lambda\hat{\sigma}_n^2).$$

Apply this, (3.14) and (3.16) in inequality (3.11) to get

$$\|I - Q_{n+1}\|_F \leq \Lambda\hat{\sigma}_n(4 + 2\Lambda\hat{\sigma}_n^2)(1 + \Lambda\hat{\sigma}_n^2) \leq (2 + r^3)(1 + r^3)r^{\frac{3}{2}}.$$

In the last line we have used the fact that by Claim 1(i), $\hat{\sigma}_n \leq r^{\frac{3}{2}}$ and $\Lambda \leq \frac{1}{2}$. Let $\gamma \geq (2 + r^3)(1 + r^3)\sqrt{r}$, then

$$\|I - Q_{n+1}\| \leq \|I - Q_{n+1}\|_F \leq \gamma r. \tag{3.20}$$

Next, we show that there is a constant α, such that

$$\|I - Q_{n+1}^{-1}\| \leq \alpha r. \tag{3.21}$$

By (3.20), $\|I - Q_{n+1}\| \leq \gamma r$. Let λ_j, $1 \leq j \leq N$ be eigenvalues of Q_{n+1}. Therefore $|1 - \lambda_j| \leq \gamma r$ for all $1 \leq j \leq N$. Also $\|Q_{n+1}^{-1}\| = \max_{1\leq j\leq N} \frac{1}{\lambda_j} \leq \frac{1}{1-\gamma r}$, assuming $\gamma r < 1$. Then

$$\|I - Q_{n+1}^{-1}\| = \|(I - Q_{n+1})Q_{n+1}^{-1}\| \leq \|I - Q_{n+1}\|\,\|Q_{n+1}^{-1}\| \leq \frac{\gamma r}{1 - \gamma r} \leq \alpha r,$$

by assuming $\frac{\gamma}{1-\gamma r} \leq \alpha$.

Claim (iii). Now we proceed to prove Claim (iii) by using induction. By definition,

$s_n = -Q_n F(x_n)$, therefore,

$$\begin{aligned}
\|F(x_{n+1})\| &= \|F(x_{n+1}) - F(x_n) - F'(x_0)s_n + F'(x_0)s_n - Q_n^{-1}s_n\| \\
&\leq \|F(x_{n+1}) - F(x_n) - F'(x_0)s_n\| + \|(I - Q_n^{-1})s_n\| \\
&\leq Lr^{\frac{3}{2}}\|s_n\| + \alpha r\|s_n\| = r(L\sqrt{r} + \alpha)\|s_n\| \leq (L\sqrt{r} + \alpha)r^{n+3}.
\end{aligned}$$

If we assume $L\sqrt{r} + \alpha \leq \zeta$, then $\|F(x_{n+1})\| \leq \zeta r^{n+3}$, as we need.

Claim (iv). Notice that from (3.20), we have $\|Q_{n+1}\| \leq 1 + \gamma r$. Define $\beta \geq 1 + \gamma r$, then

$$\|s_{n+1}\| = \| - Q_{n+1}F(x_{n+1})\| \leq \|Q_{n+1}\|\|F(x_{n+1})\| \leq \beta\zeta\ r^{n+3}.$$

Assume $\beta\zeta \leq 1$, then $\|s_{n+1}\| \leq r^{n+3}$.

Therefore by using mathematical induction we have shown Claim 1. A consequence of Claim 1(iv) is that $\{x_n\}$ is a Cauchy sequence lying in $B_r(x_0)$. Given $p, q \geq 0$,

$$\|x_p - x_{p+q}\| \leq \sum_{k=p}^{p+q-1} \|x_{k+1} - x_k\| \leq \sum_{k=p}^{p+q-1} r^{k+2} < r^2 \sum_{k=p}^{\infty} r^k = \frac{r^{p+2}}{1-r} \leq r^{p+1}.$$

Therefore $\{x_n\}$ converges to a point $x^* \in \overline{B_r(x_0)}$. By Claim 1(iii), $F(x^*) = 0$. Take $q \to \infty$ and $p = n$ in the above calculation to get $\|x_n - x^*\| \leq r^{n+1}$. Let $e_n = x_n - x^*$ then

$$\|e_n\| \leq r^{n+1}. \tag{3.22}$$

For proof of uniqueness, let $\hat{x}$ be any other root of F in $\overline{B_r(x_0)}$. Then we could show that

for any $n \geq 0$, we have $\|\hat{e}_{n+1}\| \leq r\|\hat{e}_n\|$, where $\hat{e}_n = x_n - \hat{x}$:

$$\begin{aligned}
\|\hat{e}_{n+1}\| &= \|x_{n+1} - \hat{x}\| = \|x_n + s_n - \hat{x}\| = \|x_n - Q_n F(x_n) - \hat{x}\| \\
&= \|Q_n Q_n^{-1}\, \hat{e}_n + Q_n\Big(- F(x_n) + F(\hat{x})\, \Big)\| \\
&\leq \|Q_n\| \left\|\Big(Q_n^{-1} - F'(x_0)\, \Big)\, \hat{e}_n + \Big(F'(x_0) - F'(x_n)\, \Big)\, \hat{e}_n + \Big(- F(x_n) + F(\hat{x}) + F'(x_n)\hat{e}_n\Big)\right\| \\
&\leq \|Q_n\|\, \|\hat{e}_n\| \left(\|Q_n^{-1} - I\, \| + \|F'(x_0) - F'(x_n)\| + \int_0^1 \|F'(x_n) - F'(\hat{x} + \hat{e}_n\tau)\, \|d\tau \right) \\
&\leq \|Q_n\|\, \|\hat{e}_n\| \Big(\|Q_n^{-1} - I\| + L\|x_n - x_0\| + \frac{L}{2}\|\hat{e}_n\| \Big).
\end{aligned}$$

Since $\hat{x}\,,\, x_n \in \overline{B_r(x_0)}$ then $\|\hat{e}_n\| \leq 2r$. By the above inequalities,

$$\|\hat{e}_{n+1}\| \leq \beta(\alpha + L\sqrt{r} + L)r\|\hat{e}_n\| \leq \beta(\alpha + 2L)r\|\hat{e}_n\| \leq r\|\hat{e}_n\|,$$

if we assume $\beta(\alpha + 2L) \leq 1$. Therefore $\|x_n - \hat{x}\| \leq r^{n+1}$ and consequently

$$\|\hat{x} - x^*\| \leq \|\hat{x} - x_n\| + \|x_n - x^*\| \leq 2r^{n+1}.$$

Let $n \to \infty$ to obtain the uniqueness result. We remark that it is possible to give an alternative proof of uniqueness using uniform convexity of f on $B_r(x_0)$.

Let

$$\sqrt{r} \leq \frac{1}{19}, \quad \text{and} \quad L \leq \min\{\frac{m}{2}, \frac{1}{3} - 6\sqrt{r}\}, \tag{3.23}$$

then the constants could be selected as

$$\begin{aligned}
\gamma &= 6\sqrt{r}, & \beta &= 1 + 6r\sqrt{r}, \\
\zeta &= \frac{1}{1 + 6r\sqrt{r}}, & \alpha &= 12\sqrt{r}.
\end{aligned} \tag{3.24}$$

The calculations for finding the constants are given in Appendix C. This completes the

proof of theorem. □

Theorem 3.4. Suppose that all hypotheses of Theorem 3.3 hold. Then the NCG iteration defined by (3.9) with η_n satisfying (3.18) converges superlinearly to the unique zero of F in $\overline{B_r(x_0)}$.

Proof. By Theorem 3.3 the iteration $\{x_n\}$ defined by (3.9) for $F(x)$ converges to x^*, the unique zero of F in $\overline{B_r(x_0)}$. Assume that $F(x_n) \neq 0$ for all $n \geq 0$ and so $x_n \neq x^*$. Also assume constants r and L satisfy (3.23), then for all $n \geq 0$,

$$\|I - Q_n\| \leq \gamma r, \qquad \|F(x_n)\| \leq \zeta r^{n+2}, \qquad \|Q_n\| \leq \beta,$$

$$\|I - Q_n^{-1}\| \leq \alpha r, \qquad \|s_n\| \leq r^{n+2}, \qquad \|e_n\| = \|x_n - x^*\| \leq r^{n+1}.$$

The positive constants α, γ, ζ and β are given by (3.24). Also since $x_{n+1} \in B_r(x_0)$, it follows $\|F'(x_{n+1})^{-1}\| \leq \dfrac{1}{m}$. Furthermore, it could be shown that $\|e_{n+1}\| \leq r\|e_n\|$ by using the same technique as in Theorem 3.3 to show $\|\hat{e}_{n+1}\| \leq r\|\hat{e}_n\|$ for all $n \geq 0$. The rest of the proof follows almost exactly as in Theorem 6.6 of [52]. □

Now we relax the requirement that $F'(x_0) = I$. The simple trick is to make a change of variables so that in the new coordinates, the initial Hessian is the identity. The scaled symmetric Perry NCG method is not invariant under this change, and so the iteration is modified appropriately so that the basic theory is still applicable.

More precisely, assume $F'(x_0)$ is SPD. Define $A = F'(x_0)^{-1/2}$, $\tilde{f}(\tilde{x}) = f(A\tilde{x})$ and $x_0 = A\tilde{x}_0$. Then $\tilde{F}'(\tilde{x}_0) = I$, where $\tilde{F}(\tilde{x}) = \tilde{\nabla} f(\tilde{x})$. Define the new update formula for the approximate inverse Hessian by

$$Q_{n+1} = A^2 - \frac{s_n y_n^T A^2 + A^2 y_n s_n^T}{s_n^T y_n} + \eta_n \frac{s_n s_n^T}{s_n^T y_n}, \qquad n \geq 0, \tag{3.25}$$

$Q_0 = A^2$ and suppose

$$\min\{2, \frac{2y_n^T A^2 y_n}{y_n^T s_n}\} \leq \eta_n \leq \max\{2, \frac{2y_n^T A^2 y_n}{y_n^T s_n}\}. \tag{3.26}$$

Theorem 3.5. Suppose that Assumption 3.1 holds such that $F'(x_0)$ is SPD and $\|I - F'(x_0)\| \leq cr$, where c is a nonnegative constant. With ζ, L, c dependent on r and r sufficiently small, the NCG iteration $\{x_n\}$ defined by (3.9) and (3.25) for all η_n satisfying (3.26), is well defined and converges superlinearly to a unique zero of F in $\overline{B_r(x_0)}$.

Proof. Let $A = F'(x_0)^{-1/2}$. It can easily be shown that for all $n \geq 0$, the following hold: $x_n = A\tilde{x}_n$ and $Q_n = A\tilde{Q}_n A$, where $\tilde{x}_n$ and $\tilde{Q}_n$ denote the symmetric scaled Perry NCG iterates and inverse Hessian approximations for the function $\tilde{F}(\tilde{x})$. The previous convergence theory can be applied to $\tilde{f}$. For an arbitrary matrix M define the norm

$$\|M\|^* = \|A^{-1}MA^{-1}\|_F.$$

Since $\|\cdot\|^*$ is a norm equivalent to $\|\cdot\|$, there is some positive constant c' such that $\|M\| \leq c' \|M\|^*$ for all M. Then by inequality (3.20) we have a positive constant γ such that

$$\|F'(x_0)^{-1} - Q_{n+1}\| = \|A^2 - A\tilde{Q}_{n+1}A\| \leq c'\|A^2 - A\tilde{Q}_{n+1}A\|^* = c'\|I - \tilde{Q}_{n+1}\|_F \leq c'\gamma r.$$

Furthermore

$$\begin{aligned} \|I - Q_{n+1}\| &\leq \|I - F'(x_0)^{-1}\| + \||F'(x_0)^{-1} - Q_{n+1}\| \\ &= \|F'(x_0)^{-1}\|\|I - F'(x_0)\| + \||F'(x_0)^{-1} - Q_{n+1}\| \leq (\frac{c}{m} + c'\gamma)r. \end{aligned}$$

By assuming $\hat{\gamma} = \frac{c}{m} + c'\gamma$, we get $\|I - Q_{n+1}^{-1}\| \leq \hat{\gamma}r$. Let $\lambda_j,\ 1 \leq j \leq N$ be eigenvalues of Q_{n+1}. Therefore $|1 - \lambda_j| \leq \hat{\gamma}r$ for all $1 \leq j \leq N$. Also $\|Q_{n+1}^{-1}\| = \max_{1\leq j\leq N} \left|\frac{1}{\lambda_j}\right| \leq \frac{1}{1-\hat{\gamma}r}$, assuming $\hat{\gamma}r < 1$. Then

$$\begin{aligned} \|F'(x_0) - Q_{n+1}^{-1}\| &= \|F'(x_0)(F'(x_0)^{-1} - Q_{n+1})Q_{n+1}^{-1}\| \\ &\leq \|F'(x_0)\|\,\|F'(x_0)^{-1} - Q_{n+1}\|\,\|Q_{n+1}^{-1}\| \leq \frac{M\gamma r}{1-\hat{\gamma}r} \leq \alpha r, \end{aligned}$$

by assuming $\frac{M\gamma}{1-\hat{\gamma}r} \leq \alpha$.

Proof of other claims are the same as in Theorem 3.3 and will be omitted here. Superlinear convergence comes from Theorem 3.4. □

Appendix C

This appendix provides the mathematical calculations for finding the constants in the proof of Theorem 3.3. The requirements among constants are given by:

1. $r \leq \frac{3-\sqrt{5}}{2}$,
2. $L \leq \frac{m}{2}$,
3. $\gamma \geq (2+r^3)(1+r^3)\sqrt{r}$,
4. $\gamma r < 1$ and $\alpha \geq \frac{\gamma}{1-\gamma r}$,
5. $L\sqrt{r} + \alpha \leq \zeta \leq 1$,
6. $\beta \geq 1 + \gamma r$,
7. $\beta\zeta \leq 1$,
8. $\beta(\alpha + 2L) \leq 1$.

First let $\gamma = 6\sqrt{r}$ and $\beta = 1 + \gamma r$. Therefore requirements 3 and 6 hold. Now let $\sqrt{r} \leq \frac{1}{19}$, then $\gamma r \leq \frac{1}{2}$ and with this restriction

$$\frac{\gamma}{1-\gamma r} \leq 2\gamma.$$

Let $\alpha = 2\gamma$, then requirement 4 is satisfied. Furthermore, $\gamma r \leq \frac{1}{2}$ results in $\frac{2}{3} \leq \frac{1}{1+\gamma r}$. Next choose L such that $L \leq \frac{1}{3} - 6\sqrt{r}$. Since $r \leq \frac{1}{19}$, then L stays positive and also $2L \leq \frac{2}{3} - 2\gamma$. Then

$$\alpha = 2\gamma \leq \frac{2}{3} - 2L \leq \frac{1}{1+\gamma r} - 2L = \frac{1}{\beta} - 2L,$$

and requirement 8 holds.

Now let $\zeta = \frac{1}{\beta}$, then requirement 7 holds. Notice that by the last inequality, $\zeta \geq \alpha + 2L \geq \alpha + L\sqrt{r}$, and also $\beta \geq 1$, requirement 5 holds. All together it is enough to choose $\sqrt{r} \leq \frac{1}{19}$ and

$$L \leq \min\{\frac{m}{2}, \frac{1}{3} - 6\sqrt{r}\},$$

then the constants could be selected as

$$\gamma = 6\sqrt{r}, \qquad \beta = 1 + 6r\sqrt{r},$$
$$\zeta = \frac{1}{1 + 6r\sqrt{r}}, \qquad \alpha = 12\sqrt{r}.$$

3.3 Generalized scaled memoryless BFGS method

Recall the BFGS iteration for finding a root of $F(x)$ is given by

$$\begin{aligned} x_{n+1} &= x_n + s_n, \qquad s_n = -A_n^{-1}F(x_n), \\ A_{n+1} &= A_n + \frac{y_n y_n^T}{y_n^T s_n} - \frac{A_n s_n s_n^T A_n}{s_n^T A_n s_n}\ , \end{aligned}$$

for any $n \geq 0$, where $y_n = F(x_{n+1}) - F(x_n)$, x_0 is an initial guess and A_0 is a SPD matrix. If $y_n^T s_n > 0$ in each step, the update matrix stays SPD and also by using Sherman-Morrison-Woodbury formula, the inverse of update matrix is given by

$$A_{n+1}^{-1} = A_n^{-1} + \frac{s_n s_n^T}{y_n^T s_n}(1 + \frac{y_n^T A_n^{-1} y_n}{y_n^T s_n}) - \frac{A_n^{-1} y_n s_n^T + s_n y_n^T A_n^{-1}}{y_n^T s_n}. \tag{3.27}$$

Recently scaled quasi-Newton methods have been introduced for improving the performance of the BFGS update. The purpose of these methods is to improve the condition number of the approximation to the inverse Hessian. In each iteration A_n^{-1} is replaced by $\theta_n A_n^{-1}$, where $\theta_n > 0$ is a scaling parameter. Two well-known scaling parameters are given by Oren and Luenberger [54]:

$$\theta_n^{OL} = \frac{s_n^T A_n s_n}{y_n^T s_n}, \tag{3.28}$$

and by Oren and Spedicato [55]:

$$\theta_n^{OS} = \frac{y_n^T s_n}{y_n^T A_n^{-1} y_n}. \tag{3.29}$$

In the memoryless BFGS method, the matrix A_n^{-1} in the update formula (3.27) is replaced by the identity matrix,

$$\hat{H}_{n+1} = I + \frac{s_n s_n^T}{y_n^T s_n}(1 + \frac{y_n^T y_n}{y_n^T s_n}) - \frac{y_n s_n^T + s_n y_n^T}{y_n^T s_n}. \tag{3.30}$$

The method is called memoryless in the sense that it removes the need to store an approximation of the inverse Hessian matrix at each step. We remark that the memoryless BFGS method is the scaled symmetric Perry NCG method with η_n obeying (3.8) which also satisfies (3.18). Hence our convergence theory in Section 3.2 covers the memoryless BFGS method.

If we replace A_n^{-1} by $\theta_n I$ in (3.27), the method is called scaled memoryless BFGS:

$$H_{n+1} = \theta_n I + \frac{s_n s_n^T}{y_n^T s_n}(1 + \theta_n \frac{y_n^T y_n}{y_n^T s_n}) - \theta_n \frac{y_n s_n^T + s_n y_n^T}{y_n^T s_n}. \tag{3.31}$$

The scaling parameters (3.28) and (3.29) could be written as

$$\theta_n^{OL} = \frac{s_n^T s_n}{y_n^T s_n}, \tag{3.32}$$

$$\theta_n^{OS} = \frac{y_n^T s_n}{y_n^T y_n}. \tag{3.33}$$

Define the generalized scaled memoryless BFGS method by (3.31) with

$$\theta_n = \kappa \frac{s_n^T s_n}{y_n^T s_n}, \tag{3.34}$$

for a range of values of κ in some neighbourhood of 1 to be specified later. Note that the Oren-Luenberger version of the scaled memoryless BFGS method corresponds to $\kappa = 1$. We focus on this particular value since [52] mentions that the other choice given by (3.33) yields poorer performance.

Below we give a proof of convergence of the generalized scaled memoryless BFGS method without line search, and using Kantorovich-type assumptions. With initial guess x_0 and $H_0 = I$, the generalized scaled memoryless BFGS iterations are given by

$$x_{n+1} = x_n + s_n, \qquad s_n = -H_n F(x_n), \qquad n \geq 0, \tag{3.35}$$

with the next approximation of the inverse Hessian, H_{n+1}, given by (3.31), and θ_n is the scaling parameter given by (3.34).

Theorem 3.6. Suppose that Assumption 3.1 holds and $F'(x_0) = I$. If r and L are sufficiently small (satisfy (3.55) and (3.56)), then the generalized scaled memoryless BFGS iteration $\{x_n\}$ defined by (3.35) is well defined and either $F(x_n) = 0$ for some $n \geq 0$ or $\{x_n\}$ converges to a unique zero of F in $\overline{B_r(x_0)}$.

Proof. First we present some properties of the scale parameter $\theta_n = \kappa \dfrac{s_n^T s_n}{y_n^T s_n}$. Let

$$1 - Lr^{\frac{3}{2}} \leq \kappa \leq 1 + Lr^{\frac{3}{2}}. \tag{3.36}$$

By assuming $Lr^{\frac{3}{2}} \leq \dfrac{1}{2}$, then $|1 - \kappa| \leq Lr^{\frac{3}{2}}$ and $\dfrac{1}{2} \leq \kappa \leq \dfrac{3}{2}$.
Now define $\sigma_n = x_n - x_0$, $\hat{\sigma}_n = \max\{\|\sigma_{n+1}\|, \|\sigma_n\|\}$ and $\Lambda = \dfrac{L}{m}$.
Claim 1. If $\hat{\sigma}_n \leq r^{\frac{3}{2}}$ and $\Lambda \leq \dfrac{1}{2}$, then,

$$0 < \theta_n \leq \frac{2(1 + 4\Lambda r^3)}{m}; \tag{3.37}$$

$$|1 - \theta_n| \leq 2\Lambda(1 + 4\Lambda r^3) r^{\frac{3}{2}}; \tag{3.38}$$

$$|1 - \theta_n^{-1}| \leq 4M\Lambda(1 + 4\Lambda r^3) r^{\frac{3}{2}}. \tag{3.39}$$

Notice that there is some $\tilde{x}$ between x_n and x_{n+1} so that

$$s_n^T y_n = s_n^T (F(x_{n+1}) - F(x_n)) = s_n^T F'(\tilde{x}) s_n > 0,$$

since F' is SPD in a neighborhood of the initial point. So θ_n stays positive. Also in same way as we proved (3.13), $\|y_n\| \geq m\|s_n\|$ and

$$\theta_n = \kappa \frac{s_n^T s_n}{s_n^T y_n} \leq \frac{2\|s_n\|^2}{s_n^T y_n} = 2\, \frac{\|s_n\|\|y_n\|}{s_n^T y_n}\, \frac{\|s_n\|}{\|y_n\|} \leq \frac{2}{mw}, \tag{3.40}$$

where $w \equiv \dfrac{s_n^T y_n}{\|s_n\| \, \|y_n\|} \leq 1$. Also we observe that

$$|1-\theta_n| = \frac{|y_n^T s_n - \kappa s_n^T s_n|}{y_n^T s_n} = \frac{|(y_n - \kappa s_n)^T s_n|}{y_n^T s_n} \leq \frac{\|y_n - \kappa s_n\| \|s_n\|}{y_n^T s_n} = \frac{1}{w} \frac{\|y_n - \kappa s_n\|}{\|y_n\|}. \quad (3.41)$$

Notice that by the same procedure applied in the proof of (3.12) in Theorem 3.3,

$$\|y_n - \kappa s_n\| \leq (L\hat{\sigma}_n + \left|1-\kappa\right|)\|s_n\| \leq 2Lr^{\frac{3}{2}}\|s_n\|,$$

leading to

$$\frac{\|y_n - \kappa s_n\|}{\|y_n\|} \leq 2\Lambda r^{\frac{3}{2}}. \quad (3.42)$$

Note $\Lambda \leq \dfrac{1}{2}$, now apply Lemma 2.10 and use the same calculations as we used in proving (3.15) and (3.16) to obtain

$$\frac{1}{w} \leq \frac{1}{w^2} \leq 1 + 4\Lambda r^3. \quad (3.43)$$

Substitute (3.42) and (3.43) in (3.41),

$$|1-\theta_n| \leq 2\Lambda(1+4\Lambda r^3) r^{\frac{3}{2}}. \quad (3.44)$$

Also apply (3.43) in (3.40) to obtain

$$0 < \theta_n \leq \frac{2(1+4\Lambda r^3)}{m},$$

Similarly $\|y_n\| \leq \|F'(\tilde{x})\|\|s_n\| \leq M\|s_n\|$ and therefore

$$\theta_n^{-1} = \frac{y_n^T s_n}{\kappa\|s_n\|^2} \leq \frac{2\|y_n\|}{\|s_n\|} \leq \frac{2M\|s_n\|}{\|s_n\|} = 2M. \quad (3.45)$$

So by using this and (3.44) we obtain

$$|1-\theta_n^{-1}| = \theta_n^{-1}|1-\theta_n| \leq 4M\Lambda(1+4\Lambda r^3)r^{\frac{3}{2}}. \tag{3.46}$$

Claim 2. There are some positive constants γ and α dependent on r (to be defined later), such that for $n \geq 0$,

(i) $\|x_n - x_0\| \leq r^{\frac{3}{2}}(1-r^n)$;

(ii) $\|\theta_n I - H_n\| \leq \gamma r$, H_n is invertible and $\|\theta_n^{-1} I - H_n^{-1}\| \leq \alpha r$;

(iii) $\|F(x_n)\| \leq \zeta r^{n+2}$;

(iv) $\|s_n\| \leq r^{n+2}$.

Proof of the claims is by induction on $n \geq 0$. The base case for Claim (i) is trivial. Also $\hat{\sigma}_0 \leq r^2 \leq r^{\frac{3}{2}}$. Apply Claim 1, let $\gamma \geq (1+2r^3)\sqrt{r}$ and $\alpha \geq 2M(1+2r^3)\sqrt{r}$, then the base case for Claim (ii) is true. Since $\|F(x_0)\| \leq \zeta r^2$, base case for Claim (iii) holds. Also by assuming $\zeta \leq 1$, base case for Claim (iv) holds. Next, assume all of the statements are true for some integer $n \geq 1$, we will show they hold for $n+1$.

Claim (i). Since H_n exists, so does x_{n+1} and

$$\begin{aligned} \|x_{n+1} - x_0\| &\leq \|x_{n+1} - x_n\| + \|x_n - x_0\| \leq r^{n+2} + r^{\frac{3}{2}}(1-r^n) \\ &= r^{\frac{3}{2}}(1 - r^n(1-\sqrt{r})) \leq r^{\frac{3}{2}}(1-r^{n+1}). \end{aligned}$$

In last inequality we have assumed $r \leq \dfrac{3-\sqrt{5}}{2}$, therefore $r \leq 1-\sqrt{r}$. Moreover $\|x_{n+1} - x_0\| < r$, so $x_{n+1} \in B_r(x_0)$.

Claim (ii). [5] has shown that H_{n+1} has $N-2$ eigenvalues equal to θ_n and two other eigenvalues λ^+ and λ^- satisfy $0 < \lambda^- \leq \theta_n \leq \lambda^+$, which results in H_{n+1} being SPD.

Define $P_n = I - \dfrac{s_n y_n^T}{s_n^T y_n}$, then from (3.31),

$$H_{n+1} - \theta_n I = \frac{(s_n - \theta_n y_n)s_n^T + s_n(s_n - \theta_n y_n)^T P_n^T}{s_n^T y_n}.$$

Therefore

$$\|\theta_n I - H_{n+1}\|_F \leq \frac{\|(s_n - \theta_n y_n)s_n^T\|_F}{s_n^T y_n} + \frac{\|s_n(s_n - \theta_n y_n)^T P_n^T\|_F}{s_n^T y_n}. \tag{3.47}$$

Consider the first term of this inequality,

$$\frac{\|(s_n - \theta_n y_n)s_n^T\|_F}{s_n^T y_n} = \frac{\|s_n - \theta_n y_n\|\|s_n\|}{s_n^T y_n} = \frac{1}{w}\frac{\|s_n - \theta_n y_n\|}{\|y_n\|}, \tag{3.48}$$

where $w = \dfrac{s_n^T y_n}{\|s_n\|\ \|y_n\|}$. Similarly for the second term of (3.47), use Lemma (2.8) to get

$$\frac{\|s_n(s_n - \theta_n y_n)^T P_n^T\|_F}{s_n^T y_n} \leq \frac{\|s_n\|\|s_n - \theta_n y_n\|\|P_n\|}{s_n^T y_n} = \frac{1}{w^2}\frac{\|s_n - \theta_n y_n\|}{\|y_n\|}. \tag{3.49}$$

Substitute (3.48) and (3.49) in inequality (3.47), we obtain:

$$\|\theta_n I - H_{n+1}\|_F \leq \frac{2}{w^2}\frac{\|s_n - \theta_n y_n\|}{\|y_n\|}. \tag{3.50}$$

For finding an estimation for this inequality, notice that $F'(x_0) = I$ and

$$s_n - \theta_n y_n = s_n - \theta_n\Big(F(x_{n+1}) - F(x_n)\Big) = -\theta_n\Big(F(x_{n+1}) - F(x_n) - F'(x_0)s_n\Big) + (1 - \theta_n)s_n.$$

In following we use (3.12), (3.37), (3.39) and (3.43) to obtain

$$\begin{aligned}\|s_n - \theta_n y_n\| &\leq \theta_n(L\hat{\sigma}_n + |1-\theta_n|)\|s_n\| \\ &\leq \frac{2(1+4\Lambda r^3)}{m}(Lr^{\frac{3}{2}} + 2\Lambda(1+4\Lambda r^3)r^{\frac{3}{2}})\|s_n\| \\ &\leq 2\Lambda(1+4\Lambda r^3)(1+\frac{2(1+4\Lambda r^3)}{m})r^{\frac{3}{2}}\|s_n\|.\end{aligned}$$

Therefore since $\|y_n\| \geq m\|s_n\|$,

$$\frac{\|s_n - \theta_n y_n\|}{\|y_n\|} \leq 2\Lambda\,\frac{(1+4\Lambda r^3)}{m}(1+\frac{2(1+4\Lambda r^3)}{m})r^{\frac{3}{2}}. \tag{3.51}$$

Applying this and (3.43) in (3.50), we obtain

$$\begin{aligned}\|\theta_n I - H_{n+1}\|_F &\leq \frac{2}{w^2}\frac{\|s_n - \theta_n y_n\|}{\|y_n\|} \\ &\leq 4\Lambda\,\frac{(1+4\Lambda r^3)^2}{m}(1+\frac{2(1+4\Lambda r^3)}{m})r^{\frac{3}{2}} \\ &\leq \frac{2(1+2r^3)^2}{m}(1+\frac{2(1+2r^3)}{m})r^{\frac{3}{2}}.\end{aligned}$$

Let $\gamma \geq \frac{2(1+2r^3)^2}{m}(1+\frac{2(1+2r^3)}{m})\sqrt{r}$, then

$$\|\theta_n I - H_{n+1}\| \leq \|\theta_n I - H_{n+1}\|_F \leq \gamma r. \tag{3.52}$$

Next, show that there is a constant α, such that

$$\|\theta_n^{-1} I - H_{n+1}^{-1}\| \leq \alpha r. \tag{3.53}$$

By (3.45) and (3.52) we have, $\|I - \theta_n^{-1}H_{n+1}\| \leq \frac{\gamma r}{\theta_n} \leq 2M\gamma r$. Let λ_j, $1 \leq j \leq N$ be eigenvalues of $\theta_n^{-1}H_{n+1}$. Therefore $|1-\lambda_j| \leq 2M\gamma r$. Also $\|\theta_n H_{n+1}^{-1}\| = \max_{1\leq j\leq N}\frac{1}{\lambda_j} \leq$

$\frac{1}{1-2M\gamma r}$, assuming $M\gamma r < \frac{1}{2}$. Then

$$\|H_{n+1}^{-1}\| \leq \frac{1}{\theta_n(1-2M\gamma r)} \leq \frac{2M}{1-2M\gamma r}.$$

Therefore

$$\|\theta_n^{-1}I - H_{n+1}^{-1}\| \leq \frac{\|H_{n+1}^{-1}\|}{\theta_n}\|H_{n+1} - \theta_n I\| \leq \frac{4M^2\gamma r}{1-2M\gamma r} \leq \alpha r,$$

by assuming $\frac{4M^2\gamma}{1-M\gamma r} \leq \alpha$.
Claim (iii). By definition, $s_n = -H_n F(x_n)$, then

$$\begin{aligned}
\|F(x_{n+1})\| &= \|F(x_{n+1}) - F(x_n) - H_n^{-1}s_n\| \\
&\leq \|F(x_{n+1}) - F(x_n) - F'(x_0)s_n\| + |1-\theta_n^{-1}|\|s_n\| + \|(\theta_n^{-1}I - H_n^{-1})s_n\| \\
&\leq Lr^{\frac{3}{2}}\|s_n\| + 4M\Lambda(1+4\Lambda r^3)r^{\frac{3}{2}}\|s_n\| + \alpha r\|s_n\| \\
&\leq (Lr^{\frac{3}{2}} + 2M(1+2r^3)r^{\frac{3}{2}} + \alpha r)r^{n+2} \leq (L\sqrt{r} + 4M\sqrt{r} + \alpha)r^{n+3}.
\end{aligned}$$

If we assume $L\sqrt{r} + 4M\sqrt{r} + \alpha \leq \zeta$, then $\|F(x_{n+1})\| \leq \zeta r^{n+3}$, as we need.
Claim (iv). Notice that from (3.52), we have $\|H_{n+1}\| \leq \theta_n(1+2M\gamma r) \leq \frac{2(1+2r^3)}{m}(1+2M\gamma r)$. Define $\beta \geq \frac{2(1+2r^3)}{m}(1+2M\gamma r)$, then $\|H_{n+1}\| \leq \beta$. By using definition,

$$\|s_{n+1}\| = \| - H_{n+1}F(x_{n+1})\| \leq \|H_{n+1}\|\|F(x_{n+1})\| \leq \beta\zeta\ r^{n+3}.$$

Assume $\beta\zeta \leq 1$, then $\|s_{n+1}\| \leq r^{n+3}$.

Therefore by using mathematical induction we have shown Claim 2. A consequence of Claim 2(iv) is that $\{x_n\}$ is a Cauchy sequence lying in $B_r(x_0)$. Given $p, q \geq 0$,

$$\|x_p - x_{p+q}\| \leq \sum_{k=p}^{p+q-1}\|x_{k+1} - x_k\| \leq \sum_{k=p}^{p+q-1} r^{k+2} < r^2\sum_{k=p}^{\infty} r^k = \frac{r^{p+2}}{1-r} \leq r^{p+1}.$$

Therefore $\{x_n\}$ converges to a point $x^* \in \overline{B_r(x_0)}$. According to Claim 2(iii), $F(x^*) = 0$. By taking $q \to \infty$ and $p = n$ in the above calculation, $\|x_n - x^*\| \leq r^{n+1}$. Let $e_n = x_n - x^*$ then

$$\|e_n\| \leq r^{n+1}. \tag{3.54}$$

For proof of uniqueness, let $\hat{x}$ be any other root of F in $\overline{B_r(x_0)}$. Then we could show that for any $n \geq 0$, we have $\|\hat{e}_{n+1}\| \leq r\|\hat{e}_n\|$, where $\hat{e}_n = x_n - \hat{x}$.

$$\begin{aligned}
\|\hat{e}_{n+1}\| &= \|x_{n+1} - \hat{x}\| = \|x_n + s_n - \hat{x}\| = \|x_n - H_nF(x_n) - \hat{x}\| \\
&= \|H_nH_n^{-1}\,\hat{e}_n + H_n\Big(-F(x_n) + F(\hat{x})\Big)\| \\
&\leq \Big\|H_n(H_n^{-1} - \theta_n^{-1}I)\hat{e}_n + (\theta_n^{-1} - 1)H_n\hat{e}_n\Big\| \\
&\qquad + \Big\|H_n(F'(x_0) - F'(x_n))\hat{e}_n + H_n\Big(-F(x_n) + F(\hat{x}) + F'(x_n)\hat{e}_n\Big)\Big\| \\
&\leq \|H_n\|\Big(\|H_n^{-1} - \theta_n^{-1}I\|\,\|\hat{e}_n\| + \Big|\theta_n^{-1} - 1\Big|\,\|\hat{e}_n\|\Big) \\
&\qquad + \|H_n\|\Big(\|F'(x_0) - F'(x_n)\|\,\|\hat{e}_n\| + \|-F(x_n) + F(\hat{x}) + F'(x_n)\hat{e}_n\|\Big) \\
&\leq \|H_n\|\,\|\hat{e}_n\|\Bigg(\|H_n^{-1} - \theta_n^{-1}I\| + \Big|\theta_n^{-1} - 1\Big| + \|F'(x_0) - F'(x_n)\| + \int_0^1 \|F'(x_n) - F'(\hat{x} + \hat{e}_n\tau)\,\|d\tau \\
&\leq \|H_n\|\,\|\hat{e}_n\|\Big(\alpha r + 2M(1 + 2r^3)r^{\frac{3}{2}} + Lr^{\frac{3}{2}} + \frac{L}{2}\|\hat{e}_n\|\Big).
\end{aligned}$$

Since $\hat{x}$, $x_n \in \overline{B_r(x_0)}$ then $\|\hat{e}_n\| \leq 2r$, the above inequality reduces to

$$\|\hat{e}_{n+1}\| \leq \beta\Big(\alpha + 4M\sqrt{r} + 2L\Big)r\|\hat{e}_n\|.$$

If we assume $\beta(\alpha + 4M\sqrt{r} + 2L) \leq 1$, then $\|\hat{e}_n\| \leq r^{n+1}$ and

$$\|\hat{x} - x^*\| \leq \|\hat{x} - x_n\| + \|x_n - x^*\| \leq 2r^{n+1}.$$

Let $n \to \infty$ to obtain the uniqueness result. Now assume

$$\sqrt{r} \leq \frac{m^3}{792M^2}, \tag{3.55}$$

and

$$L \leq \frac{m}{12} - \frac{64M^2}{m^2}\sqrt{r} - 2M\sqrt{r}. \tag{3.56}$$

The constants used in the proof of this theorem could be selected as,

$$\begin{aligned} \gamma &= \frac{16}{m^2}\sqrt{r}, & \beta &= \frac{4}{m}(1 + \frac{32M}{m^2}r\sqrt{r}), \\ \zeta &= \frac{m^3}{4m^2 + 128Mr\sqrt{r}}, & \alpha &= \frac{128M^2}{m^2}\sqrt{r}. \end{aligned} \tag{3.57}$$

The calculations for finding the constants are given in Appendix D. This completes proof of the theorem. □

Theorem 3.7. Suppose that all hypotheses of Theorem 3.6 hold. Then the generalized scaled memoryless BFGS iteration defined by (3.35) converges superlinearly to the unique zero of F in $\overline{B_r(x_0)}$.

Proof. By Theorem 3.6 the iterate $\{x_n\}$ defined by generalized scaled memoryless BFGS method for $F(x)$ converges to x^*, the unique zero of F in $\overline{B_r(x_0)}$. Assume that $F(x_n) \neq 0$ for all $n \geq 0$ and so $x_n \neq x^*$. Also assume constants r and L satisfy (3.55) and (3.56), then

$$\|I - H_n\| \leq \gamma r, \qquad \|F(x_n)\| \leq \zeta r^{n+2}, \qquad \|H_n\| \leq \beta,$$

$$\|I - H_n^{-1}\| \leq \alpha r, \qquad \|s_n\| \leq r^{n+2}, \qquad \|e_n\| = \|x_n - x^*\| \leq r^{n+1}.$$

The positive constants α, γ, ζ and β are given by (3.57). Also since $x_{n+1} \in B_r(x_0)$, $\|F'(x_{n+1})^{-1}\| \leq \dfrac{1}{m}$.

Furthermore, it could be shown that $\|e_{n+1}\| \leq r\|e_n\|$ by using the same technique as in the Theorem 3.6 to show $\|\hat{e}_{n+1}\| \leq r\|\hat{e}_n\|$ for all $n \geq 0$. The rest of the proof follows almost exactly as in Theorem 6.6 of [52]. □

Let us consider the case when $F'(x_0) \neq I$ but is SPD. Define $A = F'(x_0)^{-1/2}$, $\tilde{f}(\tilde{x}) = f(A\tilde{x})$ and $x_0 = A\tilde{x}_0$. Then $\tilde{F}'(\tilde{x}_0) = I$, where $\tilde{F}(\tilde{x}) = \tilde{\nabla} f(\tilde{x})$. It can be shown that the method defined by (3.35) is not invariant under this change. Define the new update formula for the approximate inverse Hessian by

$$H_{n+1} = A^2\theta_n I + \frac{s_n s_n^T}{y_n^T s_n}(1 + \theta_n \frac{y_n^T A^2 y_n}{y_n^T s_n}) - \theta_n \frac{y_n s_n^T A^2 + A^2 s_n y_n^T}{y_n^T s_n}, \qquad n \geq 0, \tag{3.58}$$

$H_0 = A^2$ and

$$\theta_n = \kappa \frac{s_n^T A^2 s_n}{y_n^T s_n}. \tag{3.59}$$

Theorem 3.8. Suppose that Assumption 3.1 holds such that $F'(x_0)$ is SPD and $\|I - F'(x_0)\| \leq cr$, where c is a nonnegative constant. With ζ, L, c dependent on r and r sufficiently small, the generalized scaled memoryless BFGS iteration $\{x_n\}$ defined by (3.35) and (3.58) for all θ_n satisfying (3.59), is well defined and converges superlinearly to a unique zero of F in $\overline{B_r(x_0)}$.

By doing some calculations it can easily be shown that for all $n \geq 0$, $x_n = A\tilde{x}_n$ and $H_n = A\tilde{H}_n A$, where $\tilde{x}_n$ and $\tilde{H}_n$ denote the generalized scaled memoryless BFGS iteration and update formula for inverse Hessian approximation for the function $\tilde{F}(\tilde{x})$. The proof of this theorem follows from that of Theorems 3.6–3.7 and will be omitted here.

Appendix D

This appendix provides the mathematical calculations for finding the constants in the proof of Theorem 3.6. The requirements among constants are given by

1. $r \leq \dfrac{3-\sqrt{5}}{2}$,
2. $Lr^{\frac{3}{2}} \leq \dfrac{1}{2}$ and $L \leq \dfrac{m}{2}$,
3. $\gamma \geq \max\{(1+2r^3)\sqrt{r}, \dfrac{2(1+2r^3)^2}{m}(1+\dfrac{2(1+2r^3)}{m})\sqrt{r}\}$,
4. $M\gamma r < \dfrac{1}{2}$ and $\alpha \geq \max\{2M(1+2r^3)\sqrt{r},\ \dfrac{4M^2\gamma}{1-2M\gamma r}\}$,
5. $\alpha + 4M\sqrt{r} + L\sqrt{r} \leq \zeta \leq 1$,
6. $\beta \geq \dfrac{2(1+2r^3)}{m}(1+2M\gamma r)$,
7. $\beta\zeta \leq 1$,
8. $\beta(\alpha + 4M\sqrt{r} + 2L) \leq 1$.

First notice that $(1+2r^3)\sqrt{r} \leq \dfrac{2(1+2r^3)^2}{m}(1+\dfrac{2(1+2r^3)}{m})\sqrt{r}$ and also $1+2r^3 \leq \sqrt{2}$ for $r \leq \dfrac{1}{2}$. Let $\gamma = \dfrac{16}{m^2}\sqrt{r}$, and notice that by assumption of theorem $m \leq 1$, therefore

$$\frac{2(1+2r^3)^2}{m}(1+\frac{2(1+2r^3)}{m})\sqrt{r} \leq \frac{4}{m}(1+\frac{2\sqrt{2}}{m})\sqrt{r} \leq \frac{16}{m^2}\sqrt{r} = \gamma.$$

Then requirement 3 holds. Let $\beta = \dfrac{4}{m}(1+2M\gamma r)$, so requirement 6 is also satisfied. Now notice that $M \geq 1$ by assumption of theorem, $2M(1+2r^3)\sqrt{r} \leq \dfrac{4M^2\gamma}{1-2M\gamma r}$. Notice that requirements 4 and 8 together give bounds for α:

$$\frac{4M^2\gamma}{1-2M\gamma r} \leq \alpha \leq \frac{m}{4(1+2M\gamma r)} - 4M\sqrt{r} - 2L.$$

Now choose r such that

$$\sqrt{r} \leq \frac{m^3}{792M^2}, \tag{3.60}$$

We could show by some calculation that this assumption implies

$$\sqrt{r} < \frac{m^3}{24(32M^2 + m^2M)}. \tag{3.61}$$

Also results in $\sqrt{r} \leq \frac{m^2}{64M}$, which gives $M\gamma r \leq \frac{1}{4}$, therefore $\frac{4M^2\gamma}{1-2M\gamma r} \leq 8M^2\gamma$ and also $\frac{m}{6} \leq \frac{m}{4(1+2M\gamma r)}$. Let $\alpha = 8M^2\gamma$ and choose L such that

$$L \leq \frac{m}{12} - \frac{64M^2}{m^2}\sqrt{r} - 2M\sqrt{r}. \tag{3.62}$$

The right-hand side of this inequality is positive by (3.61). Also from (3.62),

$$\alpha = 8M^2\gamma = \frac{128M^2}{m^2}\sqrt{r} \leq \frac{m}{6} - 4M\sqrt{r} - 2L \leq \frac{m}{4(1+2M\gamma r)} - 4M\sqrt{r} - 2L.$$

Therefore requirements 4 and 8 hold. Let $\zeta = \frac{1}{\beta}$, then requirement 7 is fulfilled. By the last inequality, $\zeta \geq \alpha + 4M\sqrt{r} + 2L \geq \alpha + 4M\sqrt{r} + L\sqrt{r}$ and also $\beta \geq 1$, so requirement 5 also holds. Notice that by assuming (3.60) requirement 1 is fulfilled and also by (3.62) requirement 2 holds since

$$L \leq \frac{m}{12} - \frac{64M^2}{m^2}\sqrt{r} - 2M\sqrt{r} \leq \frac{m}{2} \leq \frac{1}{2}.$$

All together by choosing r and L such that inequalities (3.60) and (3.62) are fulfilled, then the constants satisfying requirements 1 to 8 could be selected as,

$$\gamma = \frac{16}{m^2}\sqrt{r}, \qquad \beta = \frac{4}{m}(1 + \frac{32M}{m^2}r\sqrt{r}),$$
$$\zeta = \frac{m^3}{4m^2 + 128Mr\sqrt{r}}, \qquad \alpha = \frac{128M^2}{m^2}\sqrt{r}.$$

4

Space-time spectral collocation method

In this chapter a Chebyshev collocation method in both space and time based on the work of Tang and Xu [71] is analyzed for the heat equation. The method is shown to converge spectrally when the solution is analytic. A condition number estimate of $O(N^4)$ is derived, where N is the number of spectral modes in each direction. Also a second space-time method, which is easier to implement and has similar performance is proposed and studied. Two nonlinear PDEs, viscous Burgers' and Allen–Cahn are successfully solved numerically, hinting that these methods are also effective solvers for nonlinear PDEs.

4.1 Introduction

Spectral methods apply global smooth functions to approximate solutions of ODEs and PDEs. Its main advantage is that for analytic solutions of elliptic differential equations, the rate of convergence is an exponential function of the number of basis functions used. For problems with periodic boundary conditions, trigonometric functions can be used as basis functions, and for other boundary conditions, Legendre or Chebyshev polynomials could be used as basis functions, which are eigenfunctions of singular Sturm Liouville problems

[62]. Spectral collocation is the most popular spectral method for non-periodic functions. This method seeks an approximate solution to a given PDE (ODE) and collocates the equation at a set of interior collocation points.

For time dependent PDEs, the most common approach is to use low order finite difference approximation of the time derivative and spectral approximation of the spatial derivatives. This is not ideal since the time discretization error overwhelms the spatial discretization error. Among the first works on space-time spectral methods for PDEs with periodic boundary conditions include [68] and [67]. Other references include [63], [46], [80], [79], [69], [70] and the references therein. These works use collocation based on Gaussian quadrature in time and collocation based on Gauss-Lobatto quadrature in space.

One drawback of these space-time spectral methods is that time stepping is no longer possible. The unknowns for all times must be solved simultaneously. This presents a serious problem for PDEs in three spatial dimensions and is particularly onerous for nonlinear PDEs. It should be made clear that due to the spectral convergence, many fewer unknowns are needed compared to finite difference/element schemes for the same error tolerance. An early work on spectrally accurate ODE solvers is [37].

See [17], [26], [32], [40] and references therein for works attempting to overcome the sequential nature of time discretization schemes. Boundary value methods are stable solvers of initial value problems that achieve spectral-like convergence rates. See, for instance, [12] and [45] for two recent contributions. Spectral deferred correction methods are another class of ODE solvers with spectral convergence. See [25] and [14] for two representative publications. See [31] for a good survey of algorithms which are parallel in time. Excellent references on the theory and practice of spectral methods include [7], [9], [13], [29], [30], [35], [36], [38], [62] and [72].

In Section 4.2 we summarize the notation that will be necessary in the description of the method. In Section 4.3, we demonstrate the space-time spectral convergence of a method of Tang and Xu for the 1D heat equation. In Section 4.4, we propose a similar

space-time spectral collocation method for more general PDEs. Following this we discuss some simple iterative schemes for two nonlinear PDEs (Allen–Cahn and viscous Burgers' equations) in Section 4.5 . In Section 4.6, some numerical experiments in MATLAB are shown to confirm the theoretical results. We include an appendix which facilitates a condition number estimate of the methods.

4.2 Basic notation and preliminaries

In this thesis, let $\mathrm{M}^{\mathrm{m,n}}$ denote the space of real ($\mathbb{R}$) or complex ($\mathbb{C}$) matrices of size $m \times n$. Let a_{ij} denotes the (i,j)th entry of $A \in \mathrm{M}^{\mathrm{m,n}}$. Let I_n denotes the $n \times n$ identity matrix. For an $n \times n$ matrix M, let $[M]$ denote the $(n-1) \times (n-1)$ matrix obtained from M by deleting the last column and row, while $[\![M]\!]$ denotes the $(n-2) \times (n-2)$ matrix obtained from M by deleting the first and last columns and rows. For any complex number a, its complex conjugate is denoted by $\bar{a}$ and its real and imaginary parts are denoted by $\Re a$ and $\Im a$, respectively. For any matrix M, let M^T and M^* denote the transpose and complex conjugate transpose of M, respectively. Let $\Lambda(M)$ denote the spectrum of M. Let $|\cdot|_2$ denote the vector/matrix 2-norm and $|\cdot|_\infty$ denote the vector ∞-norm. For any vector v, denote by $\mathrm{diag}(v)$ the diagonal matrix whose diagonal entries are elements of v.

Definition 4.1. The Kronecker (tensor) product of the matrix $A \in \mathbb{C}^{p,q}$ with the matrix $B \in \mathrm{M}^{\mathrm{r,s}}$ is defined as

$$A \otimes B = \begin{bmatrix} a_{11}B & \cdots & a_{1q}B \\ \vdots & \vdots & \vdots \\ a_{p1}B & \cdots & a_{pq}B \end{bmatrix}.$$

For basic properties of Kronecker product see, for instance, [39]. Two well known and useful theorems are given below.

Theorem 4.2 ([39], Theorem 4.2.12)**.** Let $A \in \mathrm{M}^{\mathrm{m,m}}$ and $B \in \mathrm{M}^{\mathrm{n,n}}$. Furthermore, let

$\lambda \in \Lambda(A)$ with corresponding eigenvector x, and $\mu \in \Lambda(B)$ with corresponding eigenvector y. Then $\lambda\mu$ is an eigenvalue of $A \otimes B$ with corresponding eigenvector $x \otimes y$. Any eigenvalue of $A \otimes B$ arises as such a product of eigenvalues of A and B.

It follows directly that if $A \in \mathrm{M}^{\mathrm{m,m}}$ and $B \in \mathrm{M}^{\mathrm{n,n}}$ are positive (semi) definite matrices, then $A \otimes B$ is also positive (semi) definite.

Theorem 4.3 ([39], Theorem 4.4.5)**.** Let $A \in \mathrm{M}^{\mathrm{m,m}}$ and $B \in \mathrm{M}^{\mathrm{n,n}}$. Furthermore, let $\lambda \in \Lambda(A)$ with corresponding eigenvector x, and $\mu \in \Lambda(B)$ with corresponding eigenvector y. Then $\lambda + \mu$ is an eigenvalue of $(I_n \otimes A) + (B \otimes I_m)$ with corresponding eigenvector $y \otimes x$. Any eigenvalue of $(I_n \otimes A) + (B \otimes I_m)$ arises as such a sum of eigenvalues of A and B.

Definition 4.4. For any matrix $A \in \mathrm{M}^{\mathrm{m,n}}$ define

$$\mathrm{vec}(A) = (a_{11}, \cdots, a_{m1}, a_{12}, \cdots, a_{m2}, \cdots, a_{1n}, \cdots, a_{mn})^T.$$

In other words, the columns of A are stacked on top of each other to make a vector a of length mn, or A is the matrix representation of a.

The Kronecker product can be used to present linear equations in which the unknowns are matrices. One such example is:

$$AX + YB = C. \qquad \Leftrightarrow \qquad (I \otimes A)\,\mathrm{vec}(X) + (B^T \otimes I)\,\mathrm{vec}(Y) = \mathrm{vec}(C).$$

Finally, for matrices $X \in \mathrm{M}^{\mathrm{n,n}}$, $\mathrm{Y} \in \mathrm{M}^{\mathrm{m,m}}$ and $z \in \mathrm{M}^{\mathrm{mn}}$, recall that $(X \otimes Y)z = \mathrm{vec}(YZX^T)$, where $\mathrm{vec}(Z) = z$, the vector representation of Z.

Fix a positive integer N. Let P_N denote the space of polynomials of degree at most N in x for a fixed t and degree at most N in t for a fixed x. Let $x_0, \ldots, x_N$ denote the Chebyshev Gauss-Lobatto nodes with $x_0 = 1$, $x_N = -1$ and x_j descending zeros of $T_N'(x)$, where $1 \leq j \leq N-1$ and T_N is the Nth Chebyshev polynomial. The Chebyshev

Gauss-Lobatto nodes along the t axis are denoted by $\{t_j\}$. Let

$$x_h = \begin{bmatrix} x_1 \\ \vdots \\ x_{N-1} \end{bmatrix}, \qquad t_h = \begin{bmatrix} t_0 \\ \vdots \\ t_{N-1} \end{bmatrix}.$$

Note that x_h excludes both boundary points, while t_h excludes only the initial point -1. For $0 \le j \le N$, let ℓ_j be the Lagrange interpolant, a polynomial of degree N, of x_j so that $\ell_j(x_k) = \delta_{jk}$. Recall that the Chebyshev pseudospectral derivative matrix $D \in \mathbb{R}^{(N+1),(N+1)}$ has entries

$$D_{jk} = \frac{d\ell_k(x_j)}{dx}, \qquad 0 \le j, k \le N.$$

Let $d_h = D(0 : N-1, N)$, the first N entries of the last column of D. Define the Chebyshev interpolation operator as usual, for any continuous u,

$$\mathcal{I}_N u = \sum_{j=0}^{N} u(x_j)\ell_j. \tag{4.1}$$

The following is an important property of Chebyshev quadrature: for any polynomial v of degree at most $2N-1$,

$$\int_{-1}^{1} v(x)\, w(x)\, dx = \sum_{k=0}^{N} v(x_k)\rho_k, \qquad w(x) = \frac{1}{\sqrt{1-x^2}}, \tag{4.2}$$

where $\{\rho_k\}$ is the set of weights associated with Chebyshev Gauss-Lobatto quadrature. Let W_h be the $(N+1) \times (N+1)$ diagonal matrix whose diagonal entries are $\{\rho_k\}$. Denote the weighed L^2 norm of a continuous function v on $\Omega := (-1, 1)^2$ by

$$\|v\| := \left(\int_\Omega |v(x,t)|^2\, w(x)w(t)dxdt \right)^{1/2}.$$

Also, define the corresponding discrete norm

$$\|v\|_N := \left(\sum_{j,k=0}^{N} \rho_j \rho_k |v(x_j, t_k)|^2 \right)^{1/2}.$$

It is well known (inequality (5.3.2) in [13], for instance) that the weighed L^2 and discrete norms are equivalent for all polynomials v of degree at most N:

$$\|v\| \leq \|v\|_N \leq 2\|v\|. \tag{4.3}$$

Notice that throughout this chapter and next, C, c denote positive constants whose values may differ at different occurrences, but are independent of N.

4.3 Heat equation

We treat the simplest case where the spatial and temporal domains are both $(-1, 1)$. This is no loss of generality since this can always be accomplished by a simple change of variables. Consider the linear heat equation

$$u_t = u_{xx} + f(x, t), \quad \text{on} \quad (-1, 1)^2, \tag{4.4}$$

with boundary conditions $u(\pm 1, t) = 0$ and initial condition $u(x, -1) = u_0(x)$. We seek a numerical solution $u \in P_N$ at $t = 1$.

Fix an integer $N \geq 2$. We first derive the space-time Chebyshev spectral collocation method of [71]. Write

$$\ell_k(t) = \sum_{q=0}^{N} \alpha_{qk} T_q(t), \qquad 0 \leq k \leq N. \tag{4.5}$$

Let

$$c_k = \begin{cases} 2, & k = 0; \\ 1, & \text{otherwise.} \end{cases} \qquad d_k = \begin{cases} 2, & k = 0, N; \\ 1, & \text{otherwise.} \end{cases}$$

It is not difficult to show that

$$\alpha_{qk} = \begin{cases} \dfrac{2}{Nc_q d_k} \cos \dfrac{qk\pi}{N}, & 0 \le q < N; \\ \dfrac{(-1)^k}{Nd_k}, & q = N. \end{cases} \tag{4.6}$$

For any real t, define

$$u_h(t) = \begin{bmatrix} u(x_1, t) \\ \vdots \\ u(x_{N-1}, t) \end{bmatrix}, \qquad f_h(t) = \begin{bmatrix} f(x_1, t) \\ \vdots \\ f(x_{N-1}, t) \end{bmatrix}.$$

A semi-discrete approximation of the heat equation is

$$u_h'(t) = \sum_{k=0}^{N} \Big(A u_h(t_k) + f_h(t_k) \Big) \, \ell_k(t), \qquad u_h(-1) = u_{0h}, \tag{4.7}$$

where $A = [\![D^2]\!]$ and u_{0h} is the initial data evaluated at the vector of interior Chebyshev Gauss-Lobatto points x_h; i.e., $u_{0h} = u_0(x_h)$. Note that at the collocation point t_j for $0 \le j < N$,

$$u_h'(t_j) = A u_h(t_j) + f_h(t_j), \tag{4.8}$$

precisely the system of collocation equations.

Using (4.5), it follows that

$$u_h'(t) = \sum_{q,k=0}^{N} \Big(A u_h(t_k) + f_h(t_k) \Big) \, \alpha_{qk} T_q(t).$$

Integrating in time from -1 to t_j for some $0 \le j < N$, we obtain

$$u_h(t_j) - u_{0h} = \\ \sum_{k=0}^{N} \Big(A u_h(t_k) + f_h(t_k) \Big) \left[\alpha_{0k}(t_j + 1) + \frac{\alpha_{1k}(t_j^2 - 1)}{2} + \sum_{q=2}^{N} \alpha_{qk} \left(\frac{T_{q+1}(t_j)}{2(q+1)} - \frac{T_{q-1}(t_j)}{2(q-1)} - \frac{(-1)^q}{q^2 - 1} \right) \right].$$

In the above, we used the identity

$$T_q(t) = \frac{T'_{q+1}(t)}{2(q+1)} - \frac{T'_{q-1}(t)}{2(q-1)}, \qquad q \geq 2,$$

The system can be represented as

$$u_h(t_j) = A \sum_{k=0}^{N-1} B_{kj} u_h(t_k) + g_j, \qquad 0 \leq j < N, \tag{4.9}$$

where for $0 \leq k \leq N$,

$$B_{kj} = \alpha_{0k}(t_j+1) + \frac{\alpha_{1k}(t_j^2-1)}{2} + \sum_{q=2}^{N} \alpha_{qk} \left(\frac{\cos\left((q+1)\pi j/N\right)}{2(q+1)} - \frac{\cos\left((q-1)\pi j/N\right)}{2(q-1)} - \frac{(-1)^q}{q^2-1} \right),$$

and

$$g_j = \sum_{k=0}^{N} B_{kj} f_h(t_k) + B_{Nj} A u_{0h} + u_{0h}.$$

We record the following identity for future reference:

$$B_{kj} = \int_{-1}^{t_j} \ell_k(t)\, dt = \sum_{q=0}^{N} \alpha_{qk} \int_{-1}^{t_j} T_q(t)\, dt. \tag{4.10}$$

Let $V_h, G_h \in \mathbb{R}^{(N-1)\times N}$ be the matrices whose jth column is $u_h(t_j)$ and g_j, respectively, $0 \leq j < N$. Then the system becomes

$$V_h = A V_h B + G_h, \tag{4.11}$$

where $B \in \mathbb{R}^{N\times N}$ with entries $B_{kj},\ 0 \leq k, j \leq N-1$. Let $v_h = \mathrm{vec}(V_h),\ g_h = \mathrm{vec}(G_h)$ and

$$\mathcal{A} = (I_N \otimes I_{N-1}) - (B^T \otimes A).$$

Then (4.11) is equivalent to $\mathcal{A} v_h = g_h$.

We begin with some preliminary results. The first two state that B^T is a discrete

integration operator in two different senses: it exactly integrates polynomials of degree at most N evaluated at the collocation points, and its inverse differs from $[D]$ by a rank-one matrix. These facts are hardly surprising because of the way B was derived in (4.9).

Lemma 4.5. Let $N \geq 1$ and v be a complex polynomial of degree at most N so that $v(-1) = 0$. Then

$$B^T v(t_h) = \int_{-1}^{t_h} v(t)\, dt.$$

Proof. The proof is exactly the same as that for the Legendre case. See [49]. We reproduce the proof for the convenience of the reader. Since $v(-1) = 0$, we may write

$$v(t) = \sum_{k=0}^{N} v(t_k)\ell_k(t) = \sum_{k=0}^{N-1} v(t_k)\ell_k(t) = \sum_{k=0}^{N-1}\sum_{q=0}^{N} v(t_k)\alpha_{qk}T_q(t),$$

using (4.5). For $0 \leq j < N$,

$$\int_{-1}^{t_j} v(t)\, dt = \sum_{k=0}^{N-1}\sum_{q=0}^{N} v(t_k)\alpha_{qk}\int_{-1}^{t_j} T_q(t)\, dt = \sum_{k=0}^{N-1} v(t_k)B_{kj},$$

by (4.10). This proves the lemma. □

Lemma 4.6. Let $N \geq 1$. Then $B^T - [D]^{-1}$ is a rank-one matrix.

Proof. Let $\{a_k\}$ be arbitrary complex constants so that

$$u(t) = \sum_{k=0}^{N} a_k t^k, \qquad u(-1) = 0.$$

Define $u_h = u(t_h)$. Let $\mathbf{1}$ be the vector of all ones. By the above lemma,

$$\begin{aligned}
[D]B^T u_h &= [D]\int_{-1}^{t_h}\sum_{k=0}^{N-1} a_k t^k\, dt + [D]\int_{-1}^{t_h} a_N t^N\, dt \\
&= \sum_{k=0}^{N-1} a_k t_h^k + \frac{a_N}{N+1}[D]\left(t_h^{N+1} - (-1)^{N+1}\mathbf{1}\right) \\
&= u_h + a_N\left(\frac{[D]}{N+1}\left(t_h^{N+1} + (-1)^N\mathbf{1}\right) - t_h^N\right).
\end{aligned}$$

Thus $[D]B^T - I_N$ is a rank-one matrix which depends on a_N, but is independent of all other $a_j,\ 0 \le j < N$. □

The third lemma says that when applied to an analytic function evaluated at the collocation points, the quadrature error is exponentially small.

Lemma 4.7. Let $N \ge 1$ and z be analytic so that $z(-1) = 0$. Then

$$\left| B^T z(t_h) - \int_{-1}^{t_h} z(t)\, dt \right|_2 \le cN^{1/2} e^{-CN},$$

where c depends on z but is independent of N.

Proof. Let L denote the quantity on the left-hand side of the inequality of the lemma. Then

$$\begin{aligned} L &= \left| \left[B^T (\mathcal{I}_N z)(t_h) - \int_{-1}^{t_h} (\mathcal{I}_N z)(t)\, dt \right] + \int_{-1}^{t_h} (\mathcal{I}_N z - z)(t)\, dt \right|_2 \\ &\le 0 + 2\sqrt{N}\ \|\mathcal{I}_N z - z\|_{L^2(-1,1)} \\ &\le cN^{1/2} e^{-CN}. \end{aligned}$$

Note that the term inside the square brackets is zero due to Lemma 4.5, while the last inequality is a Chebyshev interpolation error estimate for analytic functions. See (5.45) in [48], for instance. □

The following results are needed to estimate the condition number of the method. The proof of the first one, one of the main technical results of this section, is postponed to the appendix.

Proposition 4.8. Let $N \ge 1$. The real part of every eigenvalue of B^T is positive.

Lemma 4.9. Let $N \ge 1$. Then $|B^T|_2 \le c$, a positive constant independent of N.

Proof. Let z_h be an unit N-vector so that $|B^T z_h|_2 = |B^T|_2$. Let z be a polynomial of degree at most N so that $z(-1) = 0$ and $z_h = z(t_h)$. By Lemma 4.5,

$$\begin{aligned}
|B^T z_h|_2^2 &= \left| \int_{-1}^{t_h} z(t)\, dt \right|_2^2 \\
&= \sum_{k=0}^{N-1} \left(\int_{-1}^{t_k} z(t) w^{1/2}(t) \cdot w^{-1/2}(t)\, dt \right)^2 \\
&\leq \sum_{k=0}^{N-1} \int_{-1}^{t_k} z^2(t)\, w(t)\, dt \int_{-1}^{1} \sqrt{1-t^2}\, dt \\
&\leq \frac{\pi N}{2} \int_{-1}^{1} z^2(t)\, w(t)\, dt \\
&\leq \frac{\pi N}{2} |[W_h^{1/2}]\, z_h|_2^2 \leq c.
\end{aligned}$$

The penultimate inequality is due to the equivalence of the discrete and weighed L^2 norms (4.3), while the last inequality follows from the fact that the weights satisfy $\rho_k \leq cN^{-1}$ for all k. □

The next lemma is well known; see inequality (7.3.5) in [13], for instance.

Lemma 4.10. Let $N \geq 2$. Then the eigenvalues of $-[\![D^2]\!]$ are real, bounded below by c and above by CN^4, where c and C are positive and independent of N.

Lemma 4.11. Let $N \geq 2$ and u be a function analytic in some open set in the complex plane containing the real interval $[-1, 1]$ and $u(\pm 1) = 0$. Let $A = [\![D^2]\!]$, where D is the Chebyshev pseudospectral derivative matrix. Then

$$|(Au(x_h) - u''(x_h))|_\infty \leq cN^3 e^{-CN}.$$

Proof. Recall the definition of the interpolation operator in (4.1). Observe that $(\mathcal{I}_N u)''(x_h) = Au(x_h)$. The result follows from the estimate ([61])

$$|(\mathcal{I}_N u - u)''(x_h)|_\infty \leq cN^3 e^{-CN}.$$

□

For any t, define the error vector

$$e_h(t) = u_h(t) - u(x_h, t),$$

where u is the solution of the heat equation (5.1) and u_h is the solution of (4.7). For $0 \leq k < N$,

$$\begin{aligned} e_h'(t_k) &= u_h'(t_k) - u_t(x_h, t_k) \\ &= Au_h(t_k) + f(x_h, t_k) - \Big(u_{xx}(x_h, t_k) + f(x_h, t_k)\Big) \\ &= Au_h(t_k) - Au(x_h, t_k) + Au(x_h, t_k) - u_{xx}(x_h, t_k) \\ &= Ae_h(t_k) + r(t_k), \end{aligned}$$

where $r(t_k) = Au(x_h, t_k) - u_{xx}(x_h, t_k)$.

Let $E_h := E_h(t_h)$ be the long vector consisting of the vectors $e_h(t_k)$ for $k = 0$ to $N-1$ stacked one on top of the other. Similarly define $\tilde{R}_h$ as the long vector consisting of vectors $r(t_k)$:

$$E_h = \begin{bmatrix} e_h(t_0) \\ \vdots \\ e_h(t_{N-1}) \end{bmatrix}, \qquad \tilde{R}_h = \begin{bmatrix} r(t_0) \\ \vdots \\ r(t_{N-1}) \end{bmatrix}. \tag{4.12}$$

The system $e_h'(t_k) = Ae_h(t_k) + r(t_k)$ for all k can be more compactly represented as

$$E_h'(t_h) = (I_N \otimes A)E_h + \tilde{R}_h.$$

Applying $B^T \otimes I_{N-1}$ on both sides leads to

$$E_h = (B^T \otimes A)E_h + (B^T \otimes I_{N-1})\tilde{R}_h + \delta,$$

where δ is a long vector with matrix representation Φ. Let Φ_j be the jth row of Φ. According to Lemma 4.7, $|\Phi_j|_2 \leq cN^{1/2}e^{-CN}$. The above equality can also be written as

$$\mathcal{A}E_h = R_h, \qquad R_h = (B^T \otimes I_{N-1})\tilde{R}_h + \delta. \tag{4.13}$$

See (4.11).

Theorem 4.12. Let $N \geq 2$ and λ be an eigenvalue of $\mathcal{A}$. Then

$$1 \leq |\lambda| \leq cN^4.$$

Proof. Let (v_h, λ) be an eigenpair of $\mathcal{A}$. Then it follows from (4.13) that $(\lambda - 1)v_h = -(B^T \otimes A)v_h$, or

$$\lambda - 1 = \gamma_j \mu_k, \tag{4.14}$$

where γ_j is some eigenvalue of B^T and μ_k is some eigenvalue of $-A$. The lower bound $|\lambda| > 1$ follows from Proposition 4.8 and Lemma 4.10.

From (4.14), an upper bound of $|\lambda|$ follows from Lemmas 4.9 and 4.10:

$$|\lambda| \leq 1 + |\gamma_j|\mu_k \leq 1 + cN^4.$$

□

We are able to derive the same upper bound of the eigenvalue magnitude using the technique of [49], but not the lower bound. The technique employed here is much simpler conceptually because the analysis reduces to an eigenvalue analysis of B^T and A.

Assume that $\mathcal{A}$ is diagonalizable so that there are diagonal G and invertible X so that $\mathcal{A} = XGX^{-1}$. Let $W = [W_h] \otimes [\![W_h]\!]$. By rescaling X if necessary, it can be assumed that $|W^{1/2}X^{-1}W^{-1/2}|_2 = 1$. For any vector x, define the norm $\nu(x) = |W^{1/2}X^{-1}x|_2$. The presence of the factor $W^{1/2}$ is so that ν scales approximately like the L^2 norm of a function

in space and time whose values are $X^{-1}x$ at the collocation points. Theorem 4.12 says that $|\lambda| \geq 1$ for any eigenvalue λ of $\mathcal{A}$. Hence $|G^{-1}|_2 \leq 1$. Since $XGX^{-1}E_h = R_h$, it follows that

$$W^{1/2}X^{-1}E_h = (W^{1/2}G^{-1}W^{-1/2})(W^{1/2}X^{-1}W^{-1/2})(W^{1/2}R_h) = G^{-1}(W^{1/2}X^{-1}W^{-1/2})(W^{1/2}R_h).$$

From Lemma 4.11, we know that if u is analytic, then $|r(t_j)|_\infty \leq cN^3e^{-CN}$ for every $0 \leq j < N$. Therefore

$$\begin{aligned}
\nu(E_h)^2 &\leq |G^{-1}|_2^2 \, |W^{1/2}X^{-1}W^{-1/2}|_2^2 \, |W^{1/2}|_2^2 \, |R_h|_2^2 \\
&\leq \frac{2c}{N}\left[|B^T|_2^2\left(\sum_{j=0}^{N-1}|r(t_j)|_2^2\right) + \sum_{j=0}^{N-1}|\Phi_j|_2^2\right] \\
&\leq \frac{2c}{N}\left[cN\left(\sum_{j=0}^{N-1}c^2N^6e^{-2CN} + \sum_{j=0}^{N-1}cNe^{-2CN}\right)\right] \\
&\leq cN^7e^{-2CN}.
\end{aligned}$$

In the above, we used Lemma 4.9 and the fact that $|M_1 \otimes M_2|_2 = |M_1|_2|M_2|_2$ for any matrices M_1, M_2.

Theorem 4.13. For any integer $N \geq 2$, let u be the solution of the heat equation (5.1). Assume that $u(x,t)$ is separately analytic in each variable. Let u_h be the solution of (4.9) and E_h be the long error vector defined in (4.12). Then

$$|W^{1/2}E_h|_2 \leq cN^{3.5}e^{-CN}.$$

Proof. In case $\mathcal{A}$ is diagonalizable, the analysis above gives

$$\nu(E_h) \leq cN^{3.5}e^{-CN}.$$

If $\mathcal{A}$ is not diagonalizable, then there is some sequence $\mathcal{A}_n$ converging to $\mathcal{A}$ so that $\mathcal{A}_n =$

$X_n G_n X_n^{-1}$ for some diagonal G_n and invertible X_n so that $|W^{1/2} X_n^{-1} W^{-1/2}|_2 = 1$. For any vector x, define

$$\nu(x) = \sup_{n \geq 1} |W^{1/2} X_n^{-1} x|_2.$$

Now proceed as before with G_n and X_n replacing G and X, respectively. Then take $n \to \infty$ to obtain, again,

$$\nu(E_h) \leq c N^{3.5} e^{-CN}.$$

The result of the theorem now follows from the first of the following inequalities:

$$c\,|W^{1/2} x|_2 \leq \nu(x) \leq |W^{1/2} x|_2, \qquad x \in \mathbb{R}^{N(N-1)}.$$

The second inequality is easy to show:

$$\nu(x) = \sup_{n \geq 1} |(W^{1/2} X_n^{-1} W^{-1/2})\,(W^{1/2} x)|_2 \leq |W^{1/2} x|_2.$$

To show the first inequality, recall that the Chebyshev weights satisfy:

$$\frac{c_1}{N} \leq \rho_j \leq \frac{c_2}{N}, \qquad 0 \leq j \leq N.$$

Let $\Lambda_{max}(M)$ and $\Lambda_{min}(M)$ denote the largest and smallest eigenvalues of M, respectively. It is not difficult to see that

$$\Lambda_{min}(W) \geq \frac{c_1^2}{N^2}, \qquad \Lambda_{max}(W) \leq \frac{c_2^2}{N^2}.$$

Let $\sigma(M)$ denote the smallest singular value of a matrix M. It is well known that

$\sigma(M_1M_2) \geq \sigma(M_1)\sigma(M_2)$. Combine above inequalities to obtain

$$\begin{aligned}
\nu(x) &\geq |(W^{1/2}X_1^{-1}W^{-1/2})\, W^{1/2}x|_2 \\
&\geq \sigma(W^{1/2}X_1^{-1}W^{-1/2})\, |W^{1/2}x|_2 \\
&\geq \sigma(W^{1/2})\sigma(W^{-1/2})\sigma(X_1^{-1})\, |W^{1/2}x|_2 \\
&\geq c\frac{\Lambda_{min}(W^{1/2})}{\Lambda_{max}(W^{1/2})}\, |W^{1/2}x|_2 \\
&\geq C\, |W^{1/2}x|_2.
\end{aligned}$$

This completes the proof of the theorem. □

We remark that for $f \in P_N$ and f_h, the long vector of f evaluated at the collocation points,

$$\left(\int_{-1}^{1}\int_{-1}^{1} |f(x,t)|^2 w(x)\,dx\; w(t)\,dt\right)^{1/2} \leq |W^{1/2}f_h|_2 \leq 2\left(\int_{-1}^{1}\int_{-1}^{1} |f(x,t)|^2 w(x)\,dx\; w(t)\,dt\right)^{1/2},$$

using the equivalence of weighed L^2 and discrete norms. This is the main reason for measuring the error in the discrete norm.

To measure the difficulty to solve a linear system with coefficient matrix M, we sometimes use the spectral condition number, defined by

$$\kappa(M) = \frac{\max_{\lambda\in\Lambda(M)} |\lambda|}{\min_{\lambda\in\Lambda(M)} |\lambda|},$$

where $\Lambda(M)$ is the spectrum of M. Using the result of Theorem 4.12, it is easy to estimate the spectral condition number of the space-time spectral collocation method.

Corollary 4.14. Let $N \geq 2$. Then

$$\kappa(\mathcal{A}) \leq cN^4.$$

Now we mention a direct solver for (4.11) based on the method of Bartels and Stew-

art [6]. Consider the Schur decompositions $A = QTQ^*$ and $B = PSP^*$, where P and Q are unitary and S and T upper triangular. Define $Y = Q^*V_hP$. A direct calculation shows that (4.11) becomes

$$T^{-1}Y - YS = T^{-1}Q^*G_hP.$$

Since T^{-1} is upper triangular, the method of Bartels and Stewart can be used to solve for Y. Note that the complexity of this method is $O(N^3)$. The algorithm of Golub, Nash and van-Loan [34] can also be used in place of that of Bartel and Stewart.

4.4 Another space-time spectral collocation method for the heat equation

We now give an alternate space-time spectral numerical method for the solution $u \in P_N$ of the heat equation (5.1). The spectral equations are

$$(I_{N+1} \otimes D)u_h = (D^2 \otimes I_{N+1})u_h + f_h,$$

where u_h and f_h are the vectors of u and f, respectively, evaluated at the collocation points. (The order of the variables is different from that of the first method for historical reasons.) Of course, since u_h vanishes at the nodes along the boundary $x = \pm 1$ and the initial value of u is known at $t = -1$, it is sufficient to solve for the unknowns $\hat{u}_h$, which is u_h deleting the components corresponding to boundary points and initial points. The resulting spectral equations are

$$(I_{N-1} \otimes [D])\hat{u}_h = (\llbracket D^2 \rrbracket \otimes I_N)\hat{u}_h + \hat{f}_h - (u_{0h} \otimes d_h),$$

where $\hat{f}_h$ is f_h removing the components corresponding to boundary points and initial points. The last term accounts for the contribution of the initial condition and is present

because the last row and column of D have been deleted. The linear equation to be solved becomes

$$A_h \hat{u}_h = \hat{f}_h - (u_{0h} \otimes d_h), \qquad A_h = (I_{N-1} \otimes [D]) - (\llbracket D^2 \rrbracket \otimes I_N). \tag{4.15}$$

Let $\mathrm{vec}(U_h) = \hat{u}_h$ and $\mathrm{vec}(F_h) = \hat{f}_h - (u_{0h} \otimes d_h)$. Here U_h and F_h are $N \times (N-1)$ matrices. Then the above equation is equivalent to the Sylvester equation $[D]U_h - U_h \llbracket D^2 \rrbracket^T = F_h$. This matrix system can be solved in $O(N^3)$ operations by the algorithm of Bartels and Stewart.

Let us see if there is any relation between this formulation and (4.11). Recall that $A = \llbracket D^2 \rrbracket$. If the two methods are equivalent, that is, they yield the same matrix equation and, of course, have the same solution (under exact arithmetic), then $V_h = U_h^T$. Taking the transpose of the second system results in $V_h[D]^T - AV_h = F_h^T$ or $V_h - AV_h[D]^{-T} = F_h^T[D]^{-T}$. Unfortunately, from Lemma 4.6, $[D]^{-T}$ is not the same as B (this can also be verified by an explicit computation for small values of N) and so the two methods are different.

In general, the second method is easier to implement for more complicated PDEs. In one sentence, the two methods differ in that the discrete heat equation is integrated analytically in time for the method of (4.11). For the PDE $u_{tt} + a(x,t)u_t = u_{xx} + f$, for instance, it is non-trivial to perform the time integration analytically. On the contrary, the code for the second method is really no more difficult than that for the wave equation.

Next we state two useful results, the first of which is a sharpening of a result proved in [65] in the context of stability theory of a linear hyperbolic PDE. The sharper result is not needed in this paper, but is crucial in the upcoming work in the next chapter. Its proof is postponed to Appendix E.

Proposition 4.15. Let $N \geq 1$. The real part of every eigenvalue of $[D]$ is positive and bounded away from zero.

Lemma 4.16. Let $N \geq 1$ and λ be an eigenvalue of $[D]$. Then $|\lambda| \leq cN^2$.

Proof. An upper bound for the magnitude of an eigenvalue of D is well known and is

also cN^2. Its proof is very similar to the proof of this lemma which is included here for completeness.

Let v_h be an eigenvector corresponding to λ and v be the unique polynomial of degree at most N so that $v(-1) = 0$ and $v(t_h) = v_h$. Note that $[D]v_h = \lambda v_h$ and $[D]v_h = v'(t_h)$, with the latter due to the fact that v is a polynomial of degree at most N and the action of $[D]$ on $v(t_h)$ gives its derivative exactly at the collocation points. It follows that $v'(t_j) = \lambda v(t_j)$ for $0 \leq j < N$ and so

$$\sum_{j=0}^{N} v'(t_j)\overline{v(t_j)}\rho_j = \lambda \sum_{j=0}^{N} |v(t_j)|^2 \rho_j.$$

Note that the above two terms corresponding to $j = N$ are both zero since $v(-1) = 0$. Since Chebyshev quadrature is exact for polynomials of degree at most $2N - 1$, the left-hand side is equal to the integral

$$\int_{-1}^{1} v'\bar{v}w \leq \sqrt{\int_{-1}^{1} |v'|^2 w} \sqrt{\int_{-1}^{1} |\bar{v}|^2 w} \leq cN^2 \int_{-1}^{1} |v|^2 w,$$

with the last inequality due to an inverse estimate (see (5.5.4) in [13], for instance). Thus

$$|\lambda| = \frac{\left|\int_{-1}^{1} v'\bar{v}w\right|}{\sum_{j=0}^{N} |v(t_j)|^2 \rho_j} \leq cN^2 \frac{\int_{-1}^{1} |v|^2 w}{\sum_{j=0}^{N} |v(t_j)|^2 \rho_j} \leq CN^2$$

by the equivalence of the discrete and weighted L^2 norms. □

The theorem below states that the spectral condition number of the discrete spectral differentiation operator scales like $O(N^4)$.

Theorem 4.17. Let $N \geq 2$. Let A_h be the Chebyshev spectral collocation matrix defined above associated with polynomials P_N. Then

$$\kappa(A_h) \leq CN^4.$$

Proof. Let $\{\gamma_j\}$ be the set of eigenvalues of $[D]$ and $\{\mu_j\}$ be those of $-[\![D^2]\!]$. From (4.15),

it follows that for some j, k,

$$\lambda = \gamma_j + \mu_k.$$

From Proposition 4.15 and Lemma 4.10, $\Re\gamma_j$, $\mu_k \geq c$ for some positive constant c independent of N. Hence $\Re\lambda \geq 2c$, which implies that $|\lambda| \geq 2c$. From Lemmas 4.16 and 4.10, it follows that

$$|\lambda| \leq CN^2 + cN^4 \leq C_1 N^4.$$

Combine the above two inequalities to obtain

$$\kappa(A_h) \leq cN^4.$$

□

The convergence analysis is very much similar to the one in the previous section. Let v be analytic in a region in the complex plane containing the real interval $[-1, 1]$ and $v(-1) = 0$. Let $0 \leq k < N$ and $\epsilon_k = (v - \mathcal{I}_N v)'(t_k)$. From [61], it is known that $|\epsilon_k|_\infty \leq cN^2 e^{-CN}$. Observe that

$$\begin{aligned} v'(t_k) &= (\mathcal{I}_N v)'(t_k) + (v - \mathcal{I}_N v)'(t_k) \\ &= \Big([D]\,(\mathcal{I}_N) v(t_h)\Big)_k + \epsilon_k \\ &= \Big([D]\, v(t_h)\Big)_k + \epsilon_k. \end{aligned}$$

Recall the error equation

$$e'(t_k) = Ae(t_k) + r(t_k), \qquad 0 \leq k < N,$$

where $r(t_k) = Au(x_h, t_k) - u_{xx}(x_h, t_k)$. Let $1 \leq j \leq N - 1$ and $e_j(t_k)$ refer to the jth

component of $e(t_k)$ and define

$$e_j(t_h) = \begin{bmatrix} e_j(t_0) \\ \vdots \\ e_j(t_{N-1}) \end{bmatrix},$$

and

$$E_h = \begin{bmatrix} e_1(t_h) \\ \vdots \\ e_{N-1}(t_h) \end{bmatrix}, \qquad \tilde{R}_h = \begin{bmatrix} r_1(t_h) \\ \vdots \\ r_{N-1}(t_h) \end{bmatrix}.$$

Note that these vectors are the same as those defined in (4.12) except for a different ordering. Then from the previous calculation, we have

$$([D]e_j(t_h))_k + \epsilon_{jk} = e_j'(t_k) = (Ae_h(t_k))_j + r_j(t_k),$$

where $|\epsilon_{jk}|_\infty \leq cN^2 e^{-CN}$, or

$$A_h E_h = R_h := \tilde{R}_h - \epsilon,$$

where ϵ is a long vector formed by stacking together vectors $\epsilon_j = [\epsilon_{j0}, \ldots, \epsilon_{jN-1}]^T$. Using exactly the same analysis as before, the following convergence result can be shown.

Theorem 4.18. For any integer $N \geq 2$, let u be the solution of the heat equation (5.1). Assume that $u(x,t)$ is separately analytic in each variable. Then

$$|W^{1/2} E_h|_2 \leq cN^{3.5} e^{-CN}.$$

4.5 Nonlinear PDEs

The purpose of this section is to show that it is very simple, in a few lines of code, to adapt the above methodology to solve some of the most common nonlinear PDEs with spectral

space-time convergence. We shall consider only two examples: Allen-Cahn equation and Burgers' equation.

4.5.1 Allen-Cahn equation

The Allen-Cahn equation is

$$u_t = u_{xx} + au(1-u^2) + f(x,t), \quad \text{on} \quad (-1,1)^2,$$

with initial condition $u(x,-1) = u_0(x)$ and homogeneous Dirichlet boundary conditions. Here a is a positive constant. A spectral scheme based on the second space-time collocation method of the previous section is

$$(I_{N+1} \otimes D)u_h = (D^2 \otimes I_{N+1})u_h + a(u_h - u_h^3) + f_h,$$

where u_h^3 is the vector whose jth component is the cube of the jth component of u_h. Deleting the known boundary and initial values, the final scheme reads

$$\Big[(I_{N-1} \otimes [D]) - (\llbracket D^2 \rrbracket \otimes I_N) - aI\Big]\, \hat{u}_h + a\hat{u}_h^3 = \hat{f}_h - (u_{0h} \otimes d_h).$$

This nonlinear system can be solved using the simple iteration ($k \geq 0$)

$$\Big[(I_{N-1} \otimes [D]) - (\llbracket D^2 \rrbracket \otimes I_N) - aI + a\,\mathrm{diag}((\hat{u}_h^{(k)})^2)\Big]\, \hat{u}_h^{(k+1)} = \hat{f}_h - (u_{0h} \otimes d_h).$$

We don't claim that this is the most efficient method or any convergence property, but that it is really very simple to implement and appears to work well.

4.5.2 Viscous Burgers' equation

Let $\epsilon > 0$. Consider the 1D viscous Burgers' equation

$$u_t + uu_x = \epsilon u_{xx} + f(x,t), \quad \text{on} \quad (-1,1)^2,$$

with boundary conditions $u_x(\pm 1, t) = 0$ and initial condition $u(x,-1) = u_0(x)$. We choose the Neumann boundary conditions to demonstrate that the space-time method also works for boundary conditions other than Dirichlet type. We seek a numerical solution $u \in P_N$ at $t = 1$. The spectral equations are

$$(I_{N+1} \otimes D)\, u_h + \frac{1}{2}(D \otimes I_{N+1})\, u_h^2 = \epsilon\, (D^2 \otimes I_{N+1})\, u_h + f_h.$$

Because the initial value of u is known at $t = -1$, it is sufficient to solve for the unknowns $\hat{u}_h$, which is u_h deleting the components corresponding to the initial points.

Let M be D^2 except that the first and last rows of M are the first and last rows of D, defined to enforce the Neumann boundary conditions. The resulting spectral equations are

$$(\tilde{I}_{N+1} \otimes [D])\, \hat{u}_h + \frac{1}{2}(\tilde{D} \otimes I_N)\, \hat{u}_h^2 = \epsilon\, (M \otimes I_N)\, \hat{u}_h + \hat{g}_h,$$

where $\hat{g}_h$ is $\hat{f}_h - (u_{0h} \otimes d_h)$ except that those entries corresponding to $x = \pm 1$ are set to zero, $\tilde{I}_{N+1}$ is I_{N+1} except the first and last diagonal entries are zeros, and $\tilde{D}$ is D except the first and last rows are replaced by a row of zeros. The nonlinear equation to be solved becomes

$$A_B \hat{u}_h + \frac{1}{2}(\tilde{D} \otimes I_N)\, \hat{u}_h^2 = \hat{g}_h, \qquad A_B = (\tilde{I}_{N+1} \times [D]) - \epsilon\, (M \otimes I_N).$$

A simple iteration to solve the above system is ($k \geq 0$)

$$\Big(A_B + \mathrm{diag}\Big((\tilde{D} \otimes I_N)\, \hat{u}_h^{(k)}\Big)\Big)\, \hat{u}_h^{(k+1)} = \hat{g}_h.$$

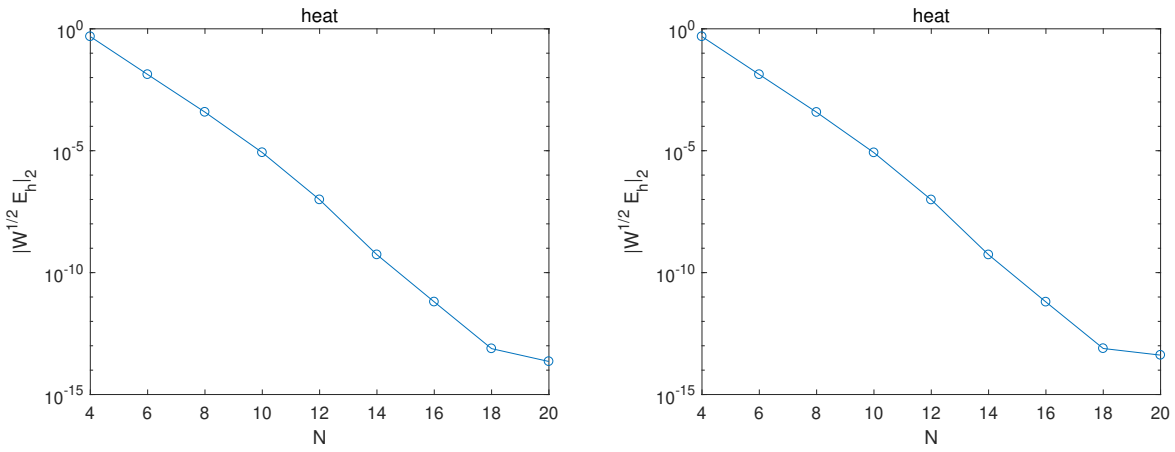

Figure 4.1: Convergence of Chebyshev collocation method (left) $\mathcal{A}$, (right) A_h for the heat equation.

See [74] for an analysis of a similar space-time spectral method for Burgers' equation.

4.6 Numerical Results

We implemented a very simple Chebyshev collocation MATLAB program. First consider the heat equation

$$u_t = u_{xx} + f,$$

with boundary conditions $u(\pm 1, t) = 0$ and initial condition $u(x, -1) = u_0(x)$. Take f so that the exact solution is $u(x,t) = e^{x+t} \sin(\pi t/2) \sin \pi x$. For the method of Tang and Xu, spectral convergence is clearly illustrated in the left figure of Figure 4.1. Note that the error E_h is $O(10^{-14})$ at $N = 18$ which corresponds to a system with 306 unknowns. The spectrum of the discrete heat operator $\mathcal{A}$ for the case $N = 60$ and a plot of the spectral condition numbers as a function of N are shown in Figure 4.2. The corresponding figures for the second method A_h are shown in the right figure of Figure 4.1 and Figure 4.3.

Now we move onto nonlinear PDEs. For both nonlinear PDEs, we take as initial guess the zero function and use the iteration defined in the previous section for each nonlinear

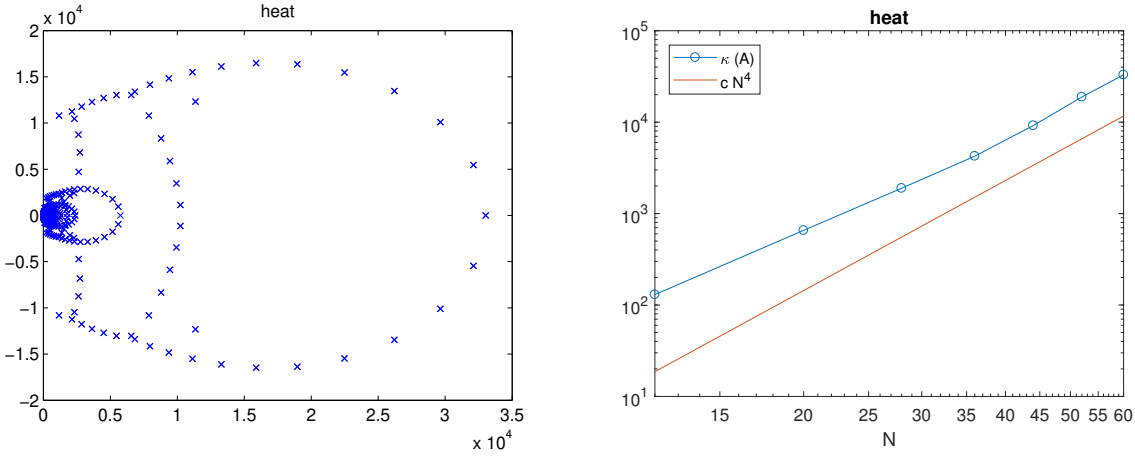

Figure 4.2: Spectrum (left) and spectral condition number (right) for the heat operator $\mathcal{A}$.

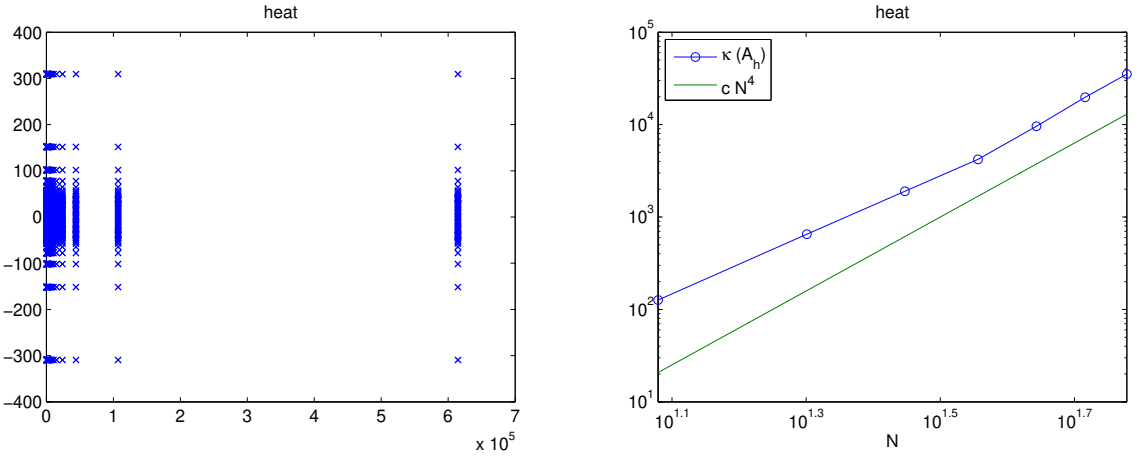

Figure 4.3: Spectrum (left) spectral condition number (right) for the heat operator A_h.

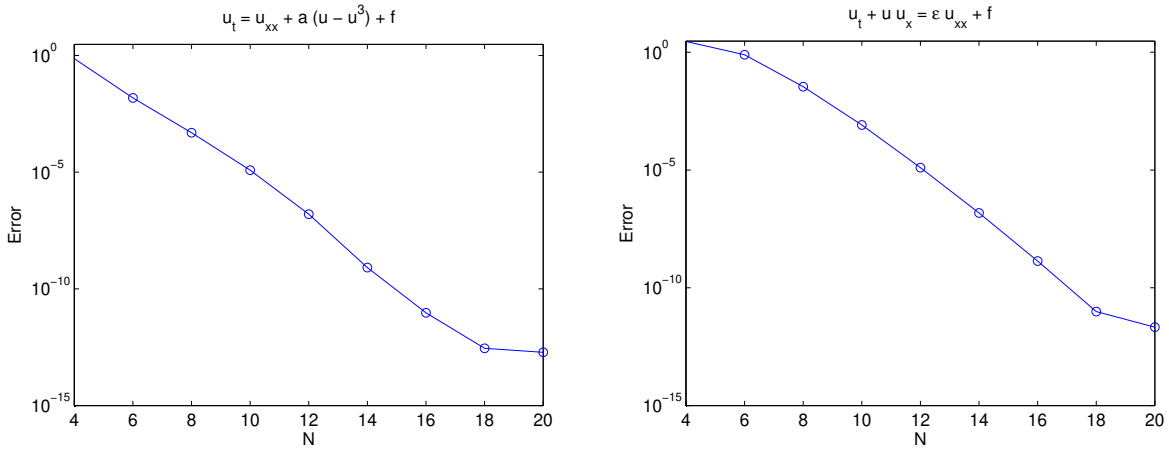

Figure 4.4: (left) Convergence of Chebyshev collocation method for the Allen-Cahn equation. (right) Convergence of Chebyshev collocation method for Burgers' equation.

PDE. Consider first the Allen-Cahn equation

$$u_t = u_{xx} + a(u - u^3) + f(x,t),$$

with homogeneous Dirichlet boundary conditions. Take $a = 0.5$ and f so that the exact solution is $u(x,t) = e^{x+t}\sin \pi x$. See the left figure of Figure 4.4 for the convergence. The stopping criterion of the iteration is that the infinity norm of the difference of two successive iterates is not more than 10^{-12}. For all values of $N \geq 8$ tested, the number of iterations decreases monotonically from 48 to 37.

Next consider the viscous Burgers' equation

$$u_t + uu_x = \epsilon u_{xx} + f(x,t),$$

with boundary conditions $u_x(\pm 1, t) = 0$ and initial condition $u(x,-1) = u_0(x)$. Take $\epsilon = 1$ and f so that the exact solution is $u(x,t) = e^t \cos(\pi x)$. With a zero initial guess, spectral convergence is clearly illustrated in Figure 4.4, the right figure. For all values of $N \geq 8$ tested, it takes between 24 and 38 iterations to solve the nonlinear system with the same stopping criterion as above.

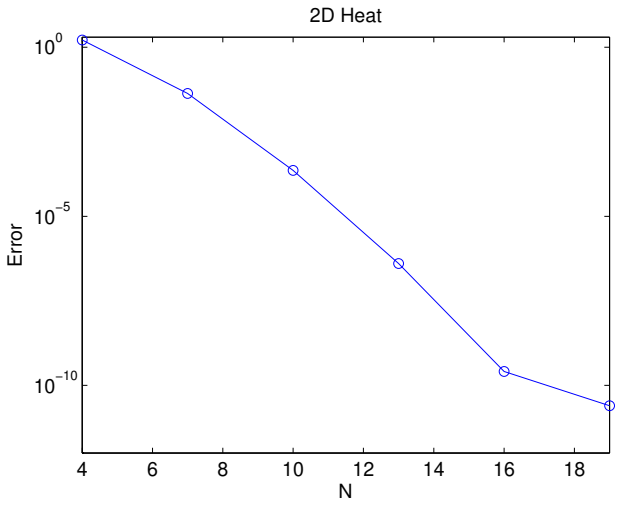

Figure 4.5: Convergence of 2D heat equation. The error is the maximum error at the final time $t = 1$.

It is straightforward to extend the methods to two spatial dimensions. As an illustration, take f so that the solution of the 2D heat equation $u_t = \Delta u + f$ on the spatial domain $(-1, 1)^2$ is

$$u(x, y) = e^{x+y+t+1} \sin \pi x \sin \pi y.$$

The convergence for the second method of Section 4.5 is given in Figure 4.5.

Appendix E

In this appendix, we prove Propositions 4.8 and 4.15.

First, the following preliminary result due to [65] is useful. Since no proof was given there, we include one here for completeness.

Lemma 4.19. Let $N \geq 1$. If $f = \sum_{k=0}^{4N-1} b_k T_k$ for some complex constants b_k, then

$$\sum_{j=0}^{N} \rho_j f(t_j) = \int_{-1}^{1} f(t)w(t)dt + \pi b_{2N}.$$

Proof. Using the definition of f,

$$\begin{aligned}
\sum_{j=0}^{N} \rho_j f(t_j) - \int_{-1}^{1} f(t)w(t)dt &= \sum_{j=0}^{N} \rho_j \sum_{k=0}^{4N-1} b_k T_k(t_j) - \int_{-1}^{1} \sum_{k=0}^{4N-1} b_k T_k(t)w(t)dt \\
&= \sum_{k=0}^{2N-1} b_k \left[\sum_{j=0}^{N} \rho_j T_k(t_j) - b_k \int_{-1}^{1} T_k(t)w(t)dt \right] \\
&\quad + b_{2N} \left[\sum_{j=0}^{N} \rho_j T_{2N}(t_j) - \int_{-1}^{1} T_{2N}(t)w(t)dt \right] \\
&\quad + \sum_{k=2N+1,\, k \text{ even}}^{4N-1} b_k \left[\sum_{j=0}^{N} \rho_j T_k(t_j) - \int_{-1}^{1} T_k(t)w(t)dt \right] \\
&\quad + \sum_{k=2N+1,\, k \text{ odd}}^{4N-1} b_k \left[\sum_{j=0}^{N} \rho_j T_k(t_j) - \int_{-1}^{1} T_k(t)w(t)dt \right] \\
&:= S_1 + S_2 + S_3 + S_4.
\end{aligned}$$

$S_1 = 0$ since Chebyshev-Gaussian quadrature is exact for any polynomial of degree at most $2N - 1$. Using the identity

$$2T_m T_n = T_{m+n} + T_{|m-n|}, \tag{4.16}$$

$T_{2N} = 2T_N^2 - 1$ follows immediately. Then for the term S_2,

$$\begin{aligned}\sum_{j=0}^{N} \rho_j T_{2N}(t_j) - \int_{-1}^{1} T_{2N}(t) w(t) dt &= \sum_{j=0}^{N} \rho_j (2T_N^2(t_j) - 1) - \int_{-1}^{1} (2T_N^2(t) - 1) w(t) dt \\ &= 2\left(\sum_{j=0}^{N} \rho_j T_N^2(t_j) - \int_{-1}^{1} T_N^2(t) w(t) dt\right) - \left(\sum_{j=0}^{N} \rho_j - \int_{-1}^{1} w(t) dt\right) \\ &= 2\left(\pi - \frac{\pi}{2}\right) - 0 = \pi.\end{aligned}$$

In the above, the definition of the Chebyshev Gauss–Lobatto points $t_j = \cos(j\pi/N)$ has been used to evaluate the penultimate sum:

$$\sum_{j=0}^{N} \rho_j T_N^2(t_j) = \sum_{j=0}^{N} \rho_j \cos^2\left(N \cos^{-1}\left[\cos\left(\frac{\pi j}{N}\right)\right]\right) = \sum_{j=0}^{N} \rho_j \cos^2(\pi j) = \sum_{j=0}^{N} \rho_j = \int_{-1}^{1} w(t) dt = \pi.$$

Therefore $S_2 = \pi b_{2N}$.

For S_3, assume $k = 2N + 2p$, for $1 \le p \le N - 1$. Then

$$\sum_{j=0}^{N} \rho_j T_{N+p}^2(t_j) = \sum_{j=0}^{N} \rho_j \cos^2\left((N+p)\frac{\pi j}{N}\right) = \sum_{j=0}^{N} \rho_j \cos^2\left(\pi j + p\frac{\pi j}{N}\right) = \int_{-1}^{1} T_p^2(t) w(t) dt.$$

From (4.16), $T_{2N+2p} = 2T_{N+p}^2 - 1$, and so

$$\begin{aligned}\sum_{j=0}^{N} \rho_j T_{2N+2p}(t_j) - \int_{-1}^{1} T_{2N+2p}(t) w(t) dt &= \sum_{j=0}^{N} \rho_j (2T_{N+p}^2(t_j) - 1) - \int_{-1}^{1} (2T_{N+p}^2(t) - 1) w(t) dt \\ &= 2\left(\sum_{j=0}^{N} \rho_j T_p^2(t_j) - \int_{-1}^{1} T_p^2(t) w(t) dt\right) - \left(\sum_{j=0}^{N} \rho_j - \int_{-1}^{1} w(t) dt\right) \\ &= 0.\end{aligned}$$

Consequently, $S_3 = 0$.

Finally assume $k = 2N + 2p + 1$ for $0 \le p \le N - 1$. Then

$$\begin{aligned}\sum_{j=0}^{N} \rho_j T_{2N+2p+1}(t_j) &= \sum_{j=0}^{N} \rho_j \cos\left((2N+2p+1)\frac{\pi j}{N}\right) \\ &= \sum_{j=0}^{N} \rho_j \cos\left(\frac{(2p+1)\pi j}{N}\right) \\ &= \int_{-1}^{1} T_{2p+1}(t) w(t) dt \;=\; 0,\end{aligned}$$

since T_{2p+1} is an odd function. By the same reason,

$$\int_{-1}^{1} T_{2N+2p+1}(t) w(t) = 0.$$

Therefore

$$\sum_{j=0}^{N} \rho_j T_{2N+2p+1}(t_j) - \int_{-1}^{1} T_{2N+2p+1}(t) w(t) dt = 0,$$

implying that $S_4 = 0$. This completes the proof. □

Next we give a proof of Proposition 4.15. As mentioned before, it is a slight improvement of a result due to [65] in a different context. The technique of proof is directly relevant to a proof of Proposition 4.8.

Proof of Proposition 4.15. Let λ be an eigenvalue of $[D]$ and v be a polynomial of degree N so that $v(-1) = 0$:

$$v = \sum_{k=0}^{N} a_k T_k, \tag{4.17}$$

where a_k are complex numbers. Suppose v satisfies the ODE

$$v'(t) = \lambda v(t) + \frac{\lambda a_N}{N}(1 - t) T'_N(t). \tag{4.18}$$

Note that the left-hand side of the above equation is a polynomial of degree $N - 1$, while

the first term on the right-hand side is a polynomial of degree N. The second term on the right-hand side is a polynomial of degree N and has a constant factor chosen so that the right-hand side is a polynomial of degree $N-1$. Observe that

$$v'(t_j) = \lambda v(t_j), \qquad 0 \le j < N. \tag{4.19}$$

When $0 < j < N$, this is because $T_N'(t_j) = 0$ by definition of the Chebyshev–Lobatto points. When $j = 0$, then $t_0 = 1$ and the equality is obvious. (4.19) is equivalent to the relation

$$[D]v(t_h) = \lambda v(t_h).$$

Using (4.17), equate the coefficient of t^{N-1} on both sides of (4.18) to obtain

$$a_{N-1} = 2a_N \left(\frac{N}{\lambda} - 1\right). \tag{4.20}$$

Note that $\lambda \neq 0$ since, otherwise, from (4.18) and the initial condition $v(-1) = 0$, it follows that $v \equiv 0$ and so $v(t_h)$ is the zero vector which cannot be an eigenvector. If $a_N = 0$, then the left-hand side of (4.18) is a polynomial of degree one less than that on the right-hand side, which is impossible. Henceforth assume $a_N \neq 0$.

Let $\beta \in (0,1)$ whose value will be determined later. Now multiply equation (4.19) by $\rho_j(1-t_j)(1+\beta t_j)\overline{v(t_j)}$ and then add the result to the complex conjugate of (4.19) multiplied by $\rho_j(1-t_j)(1+\beta t_j)v(t_j)$ and then sum to obtain

$$\sum_{j=0}^{N} \rho_j f(t_j) = 2R \sum_{j=0}^{N} \rho_j(1-t_j)(1+\beta t_j)|v(t_j)|^2 := C_1\, R, \tag{4.21}$$

where $f(t) = (1-t)(1+\beta t)(|v|^2)'(t)$ is a polynomial of degree $2N+1$ and $\lambda = R + iS$ for real R, S. Note that we can extend the above sums to $j = N$ because both terms corresponding to $j = N$ vanish. Note also that C_1 is positive since, otherwise, $v(t_j) = 0$ for $1 \le j \le N$. From (4.18), $v'(t_j) = 0$ for $1 \le j \le N$ and these conditions imply that

$v \equiv 0$. In particular $v(t_h) = 0$ and so it cannot be an eigenvector.

The goal is to show that the left-hand side of (4.21) is positive. Toward that end, write the left-hand side as $F + E$, where

$$F = \int_{-1}^{1} f(t)\, w(t)\, dt, \qquad E = \sum_{j=0}^{N} \rho_j f(t_j) - F.$$

After an integration by parts and some algebra,

$$F = \int_{-1}^{1} \frac{1 - \beta + \beta t + \beta t^2}{1 + t} \, |v(t)|^2 \, w(t)\, dt.$$

It is easy to see that $1 - \beta + \beta t + \beta t^2 \geq c$, a positive constant for $\beta \in (0, 4/5)$. Let $v(t) = z(t)\sqrt{1+t}$, where z is continuous on $[-1, 1]$ since $v(-1) = 0$. Thus

$$F \geq c \int_{-1}^{1} |z(t)|^2 w(t)\, dt =: C_2.$$

Write

$$f(t) = \sum_{k=0}^{2N+1} b_k T_k(t),$$

for some coefficients b_k. Then from Lemma 4.19, $E = \pi b_{2N}$. Thus $F + E \geq C_2 + \pi b_{2N}$.

Since the leading coefficient of T_k is 2^{k-1} and the coefficient of t^{k-1} of T_k is zero, it follows that the coefficient of t^{2N} of f is $2^{2N-1} b_{2N}$, which is equal to the coefficient of t^{2N} of the polynomial

$$-2\Re\Big(a_N \overline{a_{N-1}}\Big)(T_{N-1} T_N)' \; \beta t^2 - 2(1-\beta)|a_N|^2 T_N T_N'\, t.$$

From (4.17),

$$b_{2N} 2^{2N-1} = -2^{2N-2} \Re\Big(a_N \overline{a_{N-1}}\Big)(2N-1)\beta - 2^{2N-1}|a_N|^2 (1-\beta) N,$$

or

$$b_{2N} = -\beta\Re\Big(a_N\overline{a_{N-1}}\Big)\frac{2N-1}{2} - |a_N|^2(1-\beta)N.$$

Substitute (4.20) to get

$$b_{2N} = |a_N|^2\left(-\beta N(2N-1)\frac{R}{|\lambda|^2} + \beta(2N-1) - (1-\beta)N\right). \tag{4.22}$$

Now (4.21) becomes

$$C_1 R = E + F \geq C_2 + \pi b_{2N}.$$

Substitute (4.22) into the above to obtain

$$\left(\frac{C_1}{\pi} + \frac{|a_N|^2\beta N(2N-1)}{|\lambda|^2}\right)R \geq \frac{C_2}{\pi} + |a_N|^2\Big(\beta(3N-1) - N\Big).$$

For any value of β satisfying

$$\frac{N}{3N-1} < \beta < \frac{4}{5},$$

it is possible to deduce

$$\left(\frac{C_1}{\pi} + \frac{2\,|a_N|^2N^2}{|\lambda|^2}\right)R \geq C_3, \tag{4.23}$$

for some positive constant C_3 independent of N. It can be concluded that $R > 0$.

To show that R is bounded away from zero, first note that from (4.17), for $0 \leq k \leq N$,

$$a_k\,\|T_k\|_{0,w}^2 = \int_{-1}^{1} vT_k w, \qquad \|u\|_{0,w}^2 = \int_{-1}^{1} u^2 w,$$

leading to

$$|a_k| \leq \frac{\|v\|_{0,w}\,\|T_k\|_{0,w}}{\|T_k\|_{0,w}^2} < 1,$$

if we assume the normalization $\|v\|_{0,w} = 1$. From (4.20), it follows that

$$|a_{N-1}|^2 = 4\left|\frac{N}{\lambda} - 1\right|^2 |a_N|^2,$$

or

$$\frac{|a_N|^2 N^2}{|\lambda|^2} = \frac{|a_{N-1}|^2 N^2}{4|N-\lambda|^2} < \frac{1}{4\left|1-\frac{\lambda}{N}\right|^2} = \frac{1}{4\left[\left(1-\frac{R}{N}\right)^2 + \frac{S^2}{N^2}\right]}. \tag{4.24}$$

If $N = 1$, then $[D] = 1/2 = R$. Henceforth, assume $N \geq 2$. If $R > 1$, then we are done. Otherwise, assume $R \leq 1$. Then (4.24) becomes

$$\frac{|a_N|^2 N^2}{|\lambda|^2} < \frac{1}{4\left(1-\frac{1}{N}\right)^2} \leq 1,$$

since $N \geq 2$. Inserting this inequality in (4.23) yields immediately that $R \geq C_4$, a positive constant independent of N. □

Finally we prove the remaining proposition.

Proof of Proposition 4.8. When $N = 1, 2$, the eigenvalues of B^T are 1 and $(1 \pm i/\sqrt{3})/2$, respectively, and they have a positive real part. Henceforth, assume $N \geq 3$. Suppose $\{a_k\}$ is a set of complex constants so that

$$v = \sum_{k=0}^{N} a_k T_k, \qquad v(-1) = 0, \tag{4.25}$$

$\|v\|_{0,w} = 1$ and satisfies

$$\int_{-1}^{t} v(\tau)d\tau = \lambda v(t) + \frac{a_N}{N(N+1)}(t^2-1)T_N'(t), \tag{4.26}$$

for λ an eigenvalue of B^T. Note that the coefficient $a_N/(N(N+1))$ on the right-hand side of (4.26) has been chosen so that the coefficients of T_{N+1} on both sides of (4.26) agree. It is easy to check that $B^T v(t_h) = \lambda v(t_h)$. Using the identity

$$\int_{-1}^{t} T_k(\tau)d\tau = \frac{1}{2}\left(\frac{T_{k+1}(t)}{k+1} - \frac{T_{k-1}(t)}{k-1}\right) + \frac{(-1)^{k+1}}{k^2-1}, \qquad k \geq 2,$$

and on equating the coefficients of T_N, T_{N-1} and T_{N-2} on both sides of (4.26), we obtain

(details will be given later)

$$\begin{aligned} \frac{a_{N-1}}{2N} &= \lambda a_N, && (4.27)\\ \frac{a_{N-2}-a_N}{2(N-1)} &= \lambda a_{N-1} - \frac{a_N}{2(N+1)}, && (4.28)\\ \frac{a_{N-3}-a_{N-1}}{2(N-2)} &= \lambda a_{N-2}. && (4.29) \end{aligned}$$

(4.26) evaluated at $t = t_j$ reads

$$\int_{-1}^{t_j} v(\tau)d\tau = \lambda v(t_j), \qquad 0 \le j \le N. \tag{4.30}$$

Let $\beta \in (0,1)$ whose value will be determined later. Multiply (4.30) by $\rho_j(1-t_j)(1-\beta t_j)\overline{v(t_j)}$ and then add the result to the complex conjugate of (4.30) multiplied by $\rho_j(1-t_j)(1-\beta t_j)v(t_j)$ and then sum to obtain

$$\sum_{j=0}^{N} \rho_j f(t_j) = 2R \sum_{j=0}^{N} \rho_j (1-t_j)(1-\beta t_j)|v(t_j)|^2 := C_2\, R, \tag{4.31}$$

where

$$f(t) = (1-t)(1-\beta t)\left(\left|\int_{-1}^{t} v(\tau)d\tau\right|^2\right)',$$

is a polynomial of degree $2N+3$, $R = \Re\lambda$ and C_2 is positive. Note that each term corresponding to $j = N$ in (4.31) vanishes since $v(-1) = 0$. Write

$$f(t) = \sum_{k=0}^{2N+3} b_k T_k(t),$$

for some complex b_k. By applying Lemma 4.19,

$$\int_{-1}^{1} f(t)w(t)dt + \pi b_{2N} = C_2 R.$$

After an integration by parts, the above becomes

$$\int_{-1}^{1} \frac{1+\beta-\beta t-\beta t^2}{1+t} \left| \int_{-1}^{t} v(\tau)d\tau \right|^2 w(t)dt + \pi b_{2N} = C_2 R. \tag{4.32}$$

Use (4.25), (4.27), (4.28), (4.29) to obtain (details will be given later)

$$\begin{aligned} b_{2N} &= \frac{1}{2}\Re(\overline{a_N}a_{N-1})\left(\frac{1}{N}+\frac{1}{N+1}\right) \\ &\quad -\frac{(1+\beta)}{4}\left[\Re(\overline{a_N}a_{N-2})\left(\frac{1}{N-1}+\frac{1}{N+1}\right)+\frac{1}{N}|a_{N-1}|^2-\frac{1}{N-1}|a_N|^2\right] \\ &\quad +\frac{\beta}{8}\left[\Re(\overline{a_N}a_{N-3})\left(\frac{1}{N-2}+\frac{1}{N+1}\right)+\Re(\overline{a_{N-1}}a_{N-2})\left(\frac{1}{N-1}+\frac{1}{N}\right)\right. \\ &\qquad \left. -\Re(\overline{a_N}a_{N-1})\left(\frac{1}{N-2}+\frac{1}{N-1}\right)\right] \\ &\quad -(1+\beta)|a_N|^2\left(\frac{c_{N+1}}{(N+1)2^N}+\frac{c_N}{(N+1)2^{N-1}}\right) \\ &\quad +\beta\Re(\overline{a_N}a_{N-1})\left(\frac{c_{N+1}}{(N+1)2^{N+1}}+\frac{c_N}{N2^{N-1}}+\frac{c_{N-1}}{(N+1)2^{N-1}}\right). \end{aligned} \tag{4.33}$$

Here $c_k = -2^{N-3}N$ is the second leading coefficient of T_k:

$$T_k(t) = 2^{k-1}t^k + c_k t^{k-2} + \cdots, \qquad k \geq 2. \tag{4.34}$$

Substitute the expression for b_{2N} into (4.32) to obtain

$$\int_{-1}^{1} \frac{1+\beta-\beta t-\beta t^2}{1+t} \left| \int_{-1}^{t} v(\tau)d\tau \right|^2 w(t)dt + S_1 = (S_2+C_2)R, \tag{4.35}$$

where

$$\begin{aligned} S_1 &= \pi\frac{(1+\beta)}{4}\frac{N-1}{(N+1)^2}|a_N|^2 + \pi\frac{(1+\beta)}{4}\frac{N-1}{N(N+1)}|a_{N-1}|^2 \\ &\qquad -\pi(1+\beta)\left(\frac{c_{N+1}}{(N+1)2^N}+\frac{c_N}{(N+1)2^{N-1}}\right)|a_N|^2 > 0, \end{aligned}$$

$$\begin{aligned} S_2 &= \pi S_3 |a_N|^2 + \frac{\pi\beta}{2}\frac{(2N-1)(N-2)}{N(N+1)}|a_{N-1}|^2 \\ &\quad -2\pi\beta N|a_N|^2\left[\frac{c_{N+1}}{(N+1)2^{N+1}} + \frac{c_N}{N2^{N-1}} + \frac{c_{N-1}}{(N+1)2^{N-1}}\right], \end{aligned}$$

and

$$S_3 = -4\beta\frac{N(2N-1)(N-1)}{N+1}R^2 + 4(1+\beta)\frac{N^2}{N+1}R - \frac{3\beta}{2}\frac{1}{(N+1)^2} - \frac{2N+1}{N+1}.$$

Note that the last term of the expression for S_2 is positive since $c_k < 0$ and the coefficient of $|a_{N-1}|^2$ is positive. Note also that the integral on the left-hand side of (4.35) is positive for $\beta \in (0,1)$. Hence the remaining goal is to choose β so that S_3, a quadratic in R, is positive. Toward that end, the maximum of S_3 occurs at

$$R = \frac{1+\beta}{2\beta}\frac{N}{(2N-1)(N-1)},$$

with maximum value

$$\frac{(1+\beta)^2}{\beta}\frac{N^3}{(2N-1)(N^2-1)} - \frac{3\beta}{2}\frac{N}{(N+1)^2} - \frac{2N+1}{N+1}.$$

Hence we require

$$\frac{(1+\beta)^2}{\beta}\frac{N^3}{(2N-1)(N-1)} - \frac{3\beta}{2}\frac{N}{N+1} > 2N+1.$$

Notice that for $N \geq 3$,

$$\frac{N^2}{(2N-1)(N-1)} > \frac{1}{2}, \qquad \frac{3}{N+1} < 1.$$

Assume $\beta < 1/3$, then $1 + \beta > 4\beta$ and

$$\frac{(1+\beta)^2}{\beta}\frac{N^3}{(2N-1)(N-1)} - \frac{\beta}{2}\frac{3N}{N+1} > 16\beta N\frac{1}{2} - \frac{1}{2}\beta N = \frac{15}{2}\beta N.$$

Therefore it is enough to choose β so that

$$\frac{4N+2}{15N} < \beta < \frac{1}{3}.$$

With this choice of β, it follows from (4.35) that $R > 0$.

Next we supply some details of the above calculations. First we prove (4.27)-(4.29). Apply (4.25) in (4.26) to obtain

$$\sum_{k=0}^{N} a_k \int_{-1}^{t} T_k(\tau)d\tau = \lambda \sum_{k=0}^{N} a_k T_k(t) + \frac{a_N}{N(N+1)}(t^2-1)T_N'(t). \tag{4.36}$$

The left-hand side of this equation is

$$\begin{aligned}
\sum_{k=0}^{N} a_k \int_{-1}^{t} T_k(\tau)d\tau &= a_N \int_{-1}^{t} T_N(\tau)d\tau + a_{N-1}\int_{-1}^{t} T_{N-1}(\tau)d\tau + a_{N-2}\int_{-1}^{t} T_{N-2}(\tau)d\tau \\
&\quad + a_{N-3}\int_{-1}^{t} T_{N-3}(\tau)d\tau + \cdots \\
&= \frac{1}{2}a_N\left[\frac{T_{N+1}}{N+1} - \frac{T_{N-1}}{N-1}\right] + \frac{1}{2}a_{N-1}\left[\frac{T_N}{N} - \frac{T_{N-2}}{N-2}\right] \\
&\quad + \frac{1}{2}a_{N-2}\left[\frac{T_{N-1}}{N-1} - \frac{T_{N-3}}{N-3}\right] + \frac{1}{2}a_{N-3}\left[\frac{T_{N-2}}{N-2} - \frac{T_{N-4}}{N-4}\right] + \cdots \\
&= \frac{a_N}{2(N+1)}T_{N+1} + \frac{a_{N-1}}{2N}T_N + \left(\frac{a_{N-2}}{2(N-1)} - \frac{a_N}{2(N-1)}\right)T_{N-1} \\
&\quad + \left(\frac{a_{N-3}}{2(N-2)} - \frac{a_{N-1}}{2(N-2)}\right)T_{N-2} + \cdots .
\end{aligned} \tag{4.37}$$

The second term on the right-hand side of (4.36) is

$$\begin{aligned}
\frac{a_N}{N(N+1)}(t^2-1)T_N'(t) &= \frac{a_N}{(N+1)}(t^2-1)\left[\frac{T_N'(t)}{N}-\frac{T_{N-2}'(t)}{N-2}+\frac{T_{N-2}'(t)}{N-2}-\frac{T_{N-4}'(t)}{N-4}+\frac{T_{N-4}'(t)}{N-4}\right] \\
&= \frac{a_N}{(N+1)}2(t^2-1)\left[T_{N-1}+T_{N-3}+\frac{T_{N-4}'(t)}{2(N-4)}\right] \\
&= \frac{a_N}{(N+1)}\left[2t^2T_{N-1}+2t^2T_{N-3}-2T_{N-1}-2T_{N-3}+\cdots\right] \\
&= \frac{a_N}{(N+1)}\left[t\Big(T_N+T_{N-2}\Big)+t\Big(T_{N-2}+T_{N-4}\Big)-2T_{N-1}-2T_{N-3}+\cdots\right] \\
&= \frac{a_N}{(N+1)}\left[\frac{1}{2}\Big(T_{N+1}+T_{N-1}+T_{N-1}+T_{N-3}+T_{N-1}+T_{N-3}\right. \\
&\qquad \left.+T_{N-3}+T_{N-5}\Big)-2T_{N-1}-2T_{N-3}+\ldots\right].
\end{aligned}$$

So the right-hand side of (4.36) becomes

$$\begin{aligned}
&\lambda\sum_{k=0}^{N}a_kT_k(t)+\frac{a_N}{N(N+1)}(t^2-1)T_N'(t) \\
&\quad = \lambda a_NT_N(t)+\lambda a_{N-1}T_{N-1}(t)+\lambda a_{N-2}T_{N-2}(t)+\ldots \\
&\quad +\frac{a_N}{(N+1)}\left[\frac{1}{2}\Big(T_{N+1}+3T_{N-1}+4T_{N-3}+\cdots\Big)-2T_{N-1}-2T_{N-3}+\cdots\right] \\
&\quad = \frac{a_N}{2(N+1)}T_{N+1}+\lambda a_NT_N(t)+\left[\lambda a_{N-1}-\frac{a_N}{2(N+1)}\right]T_{N-1}(t) \\
&\qquad +\lambda a_{N-2}T_{N-2}(t)+\ldots. \qquad (4.38)
\end{aligned}$$

Therefore by equating coefficients of equations (4.37) and (4.38), we arrive at (4.27)-(4.29).

Next we show that the following equalities follow from (4.27)-(4.29):

$$\begin{aligned}
\frac{1}{2N}\Re(a_{N-1}\overline{a_N}) &= R\,|a_N|^2, && (4.39)\\
\frac{1}{2(N-1)}\Re(a_{N-2}\overline{a_{N-1}}) &= R\left(|a_{N-1}|^2+\frac{2N}{N^2-1}|a_N|^2\right), && (4.40)\\
\frac{1}{N-1}\Re(a_{N-2}\overline{a_N}) &= \left(8NR^2+\frac{2}{N^2-1}\right)|a_N|^2-\frac{1}{N}|a_{N-1}|^2, && (4.41)\\
\frac{1}{N-2}\Re(a_{N-3}\overline{a_N}) &= \left(32N(N-1)R^3+\left[\frac{4}{N+1}+\frac{2N}{N-2}\right]R\right)|a_N|^2 \\
&\qquad -\frac{6(N-1)}{N}R\,|a_{N-1}|^2. && (4.42)
\end{aligned}$$

Observe that (4.39) follows directly from (4.27). From (4.28),

$$\frac{1}{2(N-1)}a_{N-2}=\lambda a_{N-1}+\frac{1}{N^2-1}a_N.$$

Use this equation and (4.27) to arrive at (4.40) and

$$\frac{1}{N-1}\Re(a_{N-2}\overline{a_N})=\lambda a_{N-1}\overline{a_N}+\bar{\lambda}\overline{a_{N-1}}a_N+\frac{2}{N^2-1}|a_N|^2. \tag{4.43}$$

Also from (4.27)

$$\frac{1}{N}|a_{N-1}|^2=\lambda a_N\overline{a_{N-1}}+\bar{\lambda}\overline{a_N}a_{N-1}. \tag{4.44}$$

Add (4.43) and (4.44) and use (4.39) to recover (4.41). The following equations follow from (4.29) and (4.27), respectively:

$$\frac{1}{N-2}\Re(a_{N-3}\overline{a_N})=\lambda a_{N-2}\overline{a_N}+\bar{\lambda}\overline{a_{N-2}}a_N+\frac{1}{N-2}\Re(a_N\overline{a_{N-1}}), \tag{4.45}$$

and

$$\frac{1}{N}\Re(a_{N-1}\overline{a_{N-2}})=\lambda a_N\overline{a_{N-2}}+\bar{\lambda}\overline{a_N}a_{N-2}. \tag{4.46}$$

Add (4.45) and (4.46) to get

$$\frac{1}{N-2}\Re(a_{N-3}\overline{a_N}) + \frac{1}{N}\Re(a_{N-1}\overline{a_{N-2}}) = 4R\,\Re(a_N\overline{a_{N-2}}) + \frac{1}{N-2}\Re(a_N\overline{a_{N-1}}).$$

Combine the above, (4.39), (4.40) and (4.41) to get (4.42).

Finally we derive the expression for b_{2N}. Apply (4.25) and doing some calculations,

$$\begin{aligned}
\bar{v}\int v + v\int \bar{v} &= 2|a_N|^2\Big(T_N\int T_N\Big) + 2|a_{N-1}|^2\Big(T_{N-1}\int T_{N-1}\Big) \\
&\quad +2\Re(\overline{a_N}a_{N-1})\left(T_N\int T_{N-1} + T_{N-1}\int T_N\right) \\
&\quad +2\Re(\overline{a_N}a_{N-2})\left(T_N\int T_{N-2} + T_{N-2}\int T_N\right) \\
&\quad +2\Re(\overline{a_N}a_{N-3})\left(T_N\int T_{N-3} + T_{N-3}\int T_N\right) \\
&\quad +2\Re(\overline{a_{N-1}}a_{N-2})\left(T_{N-1}\int T_{N-2} + T_{N-2}\int T_{N-1}\right) + \cdots \\
&= |a_N|^2\left(T_N\left[\frac{T_{N+1}}{N+1} - \frac{T_{N-1}}{N-1}\right]\right) + |a_{N-1}|^2\left(T_{N-1}\left[\frac{T_N}{N} - \frac{T_{N-2}}{N-2}\right]\right) \\
&\quad +\Re(\overline{a_N}a_{N-1})\left(T_N\left[\frac{T_N}{N} - \frac{T_{N-2}}{N-2}\right] + T_{N-1}\left[\frac{T_{N+1}}{N+1} - \frac{T_{N-1}}{N-1}\right]\right) \\
&\quad +\Re(\overline{a_N}a_{N-2})\left(T_N\left[\frac{T_{N-1}}{N-1} - \frac{T_{N-3}}{N-3}\right] + T_{N-2}\left[\frac{T_{N+1}}{N+1} - \frac{T_{N-1}}{N-1}\right]\right) \\
&\quad +\Re(\overline{a_N}a_{N-3})\left(T_N\left[\frac{T_{N-2}}{N-2} - \frac{T_{N-4}}{N-4}\right] + T_{N-3}\left[\frac{T_{N+1}}{N+1} - \frac{T_{N-1}}{N-1}\right]\right) \\
&\quad +\Re(\overline{a_{N-1}}a_{N-2})\left(T_{N-1}\left[\frac{T_{N-1}}{N-1} - \frac{T_{N-3}}{N-3}\right] + T_{N-2}\left[\frac{T_N}{N} - \frac{T_{N-2}}{N-2}\right]\right) + \cdots,
\end{aligned}$$

where $\cdots$ denotes remaining parts in the expansion. By (4.34),

$$\begin{aligned}
&\bar{v}(t)\int_{-1}^{t} v(\tau)d\tau + v(t)\int_{-1}^{t}\bar{v}(\tau)d\tau \\
= \;& \Re(\overline{a_N}a_{N-1})\left(\frac{1}{N}+\frac{1}{N+1}\right)2^{2N-2}t^{2N} \qquad (4.47)\\
&+\left[\Re(\overline{a_N}a_{N-2})\left(\frac{1}{N-1}+\frac{1}{N+1}\right)+\frac{1}{N}|a_{N-1}|^2-\frac{1}{N-1}|a_N|^2\right]2^{2N-3}t^{2N-1}\\
&+\left[\Re(\overline{a_N}a_{N-3})\left(\frac{1}{N-2}+\frac{1}{N+1}\right)+\Re(\overline{a_{N-1}}a_{N-2})\left(\frac{1}{N-1}+\frac{1}{N}\right)\right.\\
&\quad\left.-\Re(\overline{a_N}a_{N-1})\left(\frac{1}{N-2}+\frac{1}{N-1}\right)\right]2^{2N-4}t^{2N-2}\\
&+|a_N|^2\left(\frac{c_{N+1}}{N+1}2^{N-1}+\frac{c_N}{N+1}2^N\right)t^{2N-1}\\
&+\Re(\overline{a_N}a_{N-1})(\frac{c_{N+1}}{N+1}2^{N-2}+\frac{c_N}{N}2^N+\frac{c_{N-1}}{N+1}2^N)t^{2N-2}+\cdots .
\end{aligned}$$

Then

$$\begin{aligned}
f(t) &= \left(1-(1+\beta)t+\beta t^2\right)\left(\bar{v}(t)\int_{-1}^{t} v(\tau)d\tau + v(t)\int_{-1}^{t}\bar{v}(\tau)d\tau\right)\\
&= \Re(\overline{a_N}a_{N-1})\left(\frac{1}{N}+\frac{1}{N+1}\right)2^{2N-2}t^{2N}\\
&\quad-(1+\beta)\left[\Re(\overline{a_N}a_{N-2})\left(\frac{1}{N-1}+\frac{1}{N+1}\right)+\frac{1}{N}|a_{N-1}|^2-\frac{1}{N-1}|a_N|^2\right]2^{2N-3}t^{2N}\\
&\quad+\beta\left[\Re(\overline{a_N}a_{N-3})\left(\frac{1}{N-2}+\frac{1}{N+1}\right)+\Re(\overline{a_{N-1}}a_{N-2})\left(\frac{1}{N-1}+\frac{1}{N}\right)\right.\\
&\qquad\left.-\Re(\overline{a_N}a_{N-1})\left(\frac{1}{N-2}+\frac{1}{N-1}\right)\right]2^{2N-4}t^{2N}\\
&\quad-(1+\beta)|a_N|^2\left(\frac{c_{N+1}}{N+1}2^{N-1}+\frac{c_N}{N+1}2^N\right)t^{2N}\\
&\quad+\beta\Re(\overline{a_N}a_{N-1})\left(\frac{c_{N+1}}{N+1}2^{N-2}+\frac{c_N}{N}2^N+\frac{c_{N-1}}{N+1}2^N\right)t^{2N}\\
&= 2^{2N-1}b_{2N}t^{2N}+\cdots ,
\end{aligned}$$

where b_{2N} is given by (4.33). □

5

Space-time spectral Chebyshev collocation method for linear PDEs

In this chapter a space-time Chebyshev spectral collocation method for some canonical linear PDEs including the Schrodinger, wave, Airy and beam equations is demonstrated. Fully spectral convergence as well as a condition number estimate will be given for the Schrodinger and wave equations. Numerical experiments verify the theoretical results, and further demonstrate that these methods can also solve common nonlinear PDEs such as a nonlinear reaction diffusion, Sine–Gordon, KdV, Kuramoto–Shivashinsky and Cahn–Hilliard equations.

In Section 5.1, we propose a space-time Chebyshev collocation method, the, so-called, second method in Section 4.4, for the 1D Schrodinger, wave, Airy and beam equations. We demonstrate condition number estimate of the method for Schrodinger and wave equations in Section 5.2. We discuss Spectral convergence of the proposed method in Section 5.3. In Section 5.4, we briefly discuss some simple iterative schemes for some common nonlinear PDEs. Numerical experiments in MATLAB are shown in Section 5.5, confirming the theoretical results.

5.1 Linear PDEs

In Section 4.4 we considered the linear heat equation

$$u_t = u_{xx} + F(x,t) \text{ on } (-1,1)^2,$$

with boundary conditions $u(\pm 1, t) = 0$ and initial condition $u(x,-1) = u_0(x)$. The following space-time Chebyshev collocation method was proposed,

$$(I_{N+1} \otimes D)u_h = (D^2 \otimes I_{N+1})u_h + f_h,$$

where

$$u_h = \begin{bmatrix} u(x_h, t_0) \\ \vdots \\ u(x_h, t_N) \end{bmatrix}, \qquad f_h = \begin{bmatrix} F(x_h, t_0) \\ \vdots \\ F(x_h, t_N) \end{bmatrix}.$$

Of course, since u_h vanishes at the boundary $x = \pm 1$ and the initial value of u is known at $t = -1$, it is sufficient to solve for the unknowns $\hat{u}_h$, which is u_h deleting the components corresponding to boundary points and initial points. The resulting spectral equations are

$$(I_{N-1} \otimes [D])\hat{u}_h = (\llbracket D^2 \rrbracket \otimes I_N)\hat{u}_h + \hat{f}_h - (u_{0h} \otimes d_h),$$

where u_{0h} is u_0 evaluated at the (interior spatial) collocation points and $\hat{f}_h$ is f_h removing the components corresponding to boundary points and initial points. The last term accounts for the contribution of the initial condition and is present because the last row and column of D have been deleted. The linear equation to be solved becomes

$$A_h \hat{u}_h = \hat{f}_h - (u_{0h} \otimes d_h), \qquad A_h = (I_{N-1} \times [D]) - (\llbracket D^2 \rrbracket \otimes I_N).$$

In the following, we consider other common linear PDEs in applications: Schrodinger,

wave, Airy and beam equations. We treat the simplest case where the spatial and temporal domains are both $(-1,1)$. This is no loss of generality since this can always be accomplished by a simple change of variables.

5.1.1 Schrodinger equation

The linear Schrodinger equation is

$$u_t = iu_{xx} + F(x,t), \quad \text{on } (-1,1)^2,$$

with boundary conditions $u(\pm 1,t) = 0$ and initial condition $u(x,-1) = u_0(x)$. Here $i = \sqrt{-1}$. We seek a numerical solution in P_N at $t = 1$. The spectral equations are

$$A_s \hat{u}_h = \hat{f}_h - (u_{0h} \otimes d_h), \qquad A_s = (I_{N-1} \otimes [D]) - i\,([\![D^2]\!] \otimes I_N), \tag{5.1}$$

which are very similar to those for the heat equation.

5.1.2 Wave equation

Consider the linear wave equation

$$u_{tt} = u_{xx} + F(x,t), \quad \text{on } (-1,1)^2,$$

with boundary conditions $u(\pm 1,t) = 0$ and initial conditions $u(x,-1) = u_0(x)$ and $u_t(x,-1) = u_1(x)$. We seek a numerical solution in P_N at $t = 1$. First write the PDE as a first order system for $v = [v_1, v_2]^T := [u, u_t]^T$

$$v_t = \begin{bmatrix} 0 & I \\ \partial_{xx} & 0 \end{bmatrix} v + \begin{bmatrix} 0 \\ F \end{bmatrix}, \qquad v(\pm 1,t) = 0,\ v(x,-1) = \begin{bmatrix} u_0(x) \\ u_1(x) \end{bmatrix}.$$

For $j = 1, 2$, let v_{jh} be the vector of v_j evaluated at the collocation points. The spectral equations in matrix form are

$$\begin{bmatrix} I_{N+1} \otimes D & 0 \\ 0 & I_{N+1} \otimes D \end{bmatrix} \begin{bmatrix} v_{1h} \\ v_{2h} \end{bmatrix} = \begin{bmatrix} 0 & I_{N+1} \otimes I_{N+1} \\ D^2 \otimes I_{N+1} & 0 \end{bmatrix} \begin{bmatrix} v_{1h} \\ v_{2h} \end{bmatrix} + \begin{bmatrix} 0 \\ f_h \end{bmatrix},$$

where f_h is F evaluated at the collocation points. Again, since the solution vanishes at the boundary and the initial values are known, it is only necessary to solve for a subset of those values. Using the $\hat{}$ notation to denote vectors stripping away those corresponding to boundary and initial points, the spectral equations are

$$\begin{bmatrix} I_{N-1} \otimes [D] & 0 \\ 0 & I_{N-1} \otimes [D] \end{bmatrix} \begin{bmatrix} \hat{v}_{1h} \\ \hat{v}_{2h} \end{bmatrix} = \begin{bmatrix} 0 & I_{N-1} \otimes I_N \\ [\![D^2]\!] \otimes I_N & 0 \end{bmatrix} \begin{bmatrix} \hat{v}_{1h} \\ \hat{v}_{2h} \end{bmatrix} + \begin{bmatrix} 0 \\ \hat{f}_h \end{bmatrix} - \begin{bmatrix} u_{0h} \otimes d_h \\ u_{1h} \otimes d_h \end{bmatrix},$$

where u_{0h} and u_{1h} are u_0 and u_1 evaluated at the interior collocation points. From the first equation, it follows that

$$\hat{v}_{2h} = (I_{N-1} \otimes [D])\, \hat{v}_{1h} + (u_{0h} \otimes d_h).$$

Substitute this into the second equation to get, after some algebra, the final spectral equation

$$A_w \hat{v}_{1h} = \hat{f}_h - \Big(u_{0h} \otimes ([D]\, d_h)\Big) - (u_{1h} \otimes d_h), \qquad A_w = (I_{N-1} \otimes [D]^2) - ([\![D^2]\!] \otimes I_N). \quad (5.2)$$

5.1.3 Airy equation

Consider the Airy equation

$$u_t + u_{xxx} = F(x,t), \quad \text{on } (-1,1)^2,$$

with boundary conditions $u(\pm 1, t) = 0 = u_x(1, t)$ and initial condition $u(x, -1) = u_0(x)$. We seek a numerical solution in P_N at $t = 1$. The spectral equations are

$$(I_{N+1} \otimes D)u_h + (D^3 \otimes I_{N+1})u_h = f_h,$$

where f_h is the vector of F evaluated at the (spatial and temporal) collocation points. Of course, since u_h vanishes at the boundary $x = \pm 1$ and the initial value of u is known at $t = -1$, it is sufficient to solve for the unknowns $\hat{u}_h$, which is u_h deleting the components corresponding to boundary points and initial points. Let us define the spectral approximation of the third derivative, taking into account the boundary conditions.

Let $Y = Y(x)$ be a polynomial so that $Y(\pm 1) = 0 = Y'(1)$. Let Z vanish at ± 1 so that $Y(x) = (1 - x)Z(x)$. Note that Y clearly satisfies all the boundary conditions. A simple calculation leads to

$$Y'''(x) = (1 - x)Z'''(x) - 3Z''(x). \tag{5.3}$$

It should be clear now that a spectral approximation of the third derivative satisfying the three boundary conditions is

$$B := (C \, [\![D^3]\!] - 3 \, [\![D^2]\!]) \, C^{-1}, \tag{5.4}$$

where C is an $(N - 1) \times (N - 1)$ diagonal matrix whose diagonal entries are $1 - x_j$, $1 \leq j \leq N - 1$. The resulting spectral equations are

$$(I_{N-1} \otimes [D])\hat{u}_h + (B \otimes I_N)\hat{u}_h = \hat{f}_h - (u_{0h} \otimes d_h),$$

where u_{0h} is u_0 evaluated at the (interior spatial) collocation points and $\hat{f}_h$ is f_h removing the components corresponding to boundary points and initial points. The linear equation

to be solved becomes

$$A_a \hat{u}_h = \hat{f}_h - (u_{0h} \otimes d_h), \qquad A_a = (I_{N-1} \otimes [D]) + (B \otimes I_N). \tag{5.5}$$

5.1.4 Beam equation

Finally, consider the fourth order beam equation

$$u_{tt} + u_{xxxx} = F(x,t), \quad \text{on } (-1,1)^2,$$

with boundary conditions $u(\pm 1, t) = 0 = u_x(\pm 1, t)$ and initial conditions $u(x,-1) = u_0(x)$ and $u_t(x,-1) = u_1(x)$. We seek a numerical solution in P_N at $t = 1$. As before, we write the PDE as a first order system for $v = [v_1, v_2]^T := [u, u_t]^T$,

$$v_t = \begin{bmatrix} 0 & I \\ -\partial_{xxxx} & 0 \end{bmatrix} v + \begin{bmatrix} 0 \\ F \end{bmatrix}, \qquad v(\pm 1, t) = 0, \; v(x,-1) = \begin{bmatrix} u_0(x) \\ u_1(x) \end{bmatrix}.$$

Using the same notation as before, the spectral equations in matrix form are

$$\begin{bmatrix} I_{N+1} \otimes D & 0 \\ 0 & I_{N+1} \otimes D \end{bmatrix} \begin{bmatrix} v_{1h} \\ v_{2h} \end{bmatrix} = \begin{bmatrix} 0 & I_{N+1} \otimes I_{N+1} \\ -D^4 \otimes I_{N+1} & 0 \end{bmatrix} \begin{bmatrix} v_{1h} \\ v_{2h} \end{bmatrix} + \begin{bmatrix} 0 \\ f_h \end{bmatrix}.$$

Again, the components of v_{jh} along the boundary and initial points must be removed. However, it is not as simple as before since Neumann boundary conditions must also be imposed. There are at least three ways to do this. One is to impose the boundary conditions explicitly as constraints, as in spectral tau methods. A second approach to approximate the fourth derivative is to write $Y(x) = (1-x^2)Z(x)$, so that Y automatically satisfies the boundary conditions if Z vanishes at the boundary. Then

$$Y''''(x) = (1-x^2)Z''''(x) - 8xZ'''(x) - 12Z''(x). \tag{5.6}$$

The spectral approximation of the fourth derivative, taking into account of the boundary conditions, is

$$B := \left(C\,[\![D^4]\!] - 8X\,[\![D^3]\!] - 12\,[\![D^2]\!]\right) C^{-1}, \tag{5.7}$$

where C and X are $(N-1)\times(N-1)$ diagonal matrices with diagonal entries $1-x_j^2$ and x_j, respectively. See, for instance, [72]. Another approach, suggested in [48], gives a symmetric matrix approximation of the fourth derivative accommodating the boundary conditions. There is no particular advantage in the current application since the discrete time derivative is not symmetric. This last approach appears to only work for Legendre collocation and not for Chebyshev collocation. For these reasons, we apply the second approach.

The spectral equations for $\hat{v}_{jh}$, which is v_{jh} removing the variables corresponding to boundary and initial points, become

$$\begin{bmatrix} I_{N-1}\otimes[D] & 0 \\ 0 & I_{N-1}\otimes[D] \end{bmatrix} \begin{bmatrix} \hat{v}_{1h} \\ \hat{v}_{2h} \end{bmatrix} = \begin{bmatrix} 0 & I_{N-1}\otimes I_N \\ -B\otimes I_N & 0 \end{bmatrix} \begin{bmatrix} \hat{v}_{1h} \\ \hat{v}_{2h} \end{bmatrix} + \begin{bmatrix} 0 \\ f_h \end{bmatrix} - \begin{bmatrix} u_{0h}\otimes d_h \\ u_{1h}\otimes d_h \end{bmatrix}.$$

From the first equation, it follows that

$$\hat{v}_{2h} = \left(I_{N-1}\otimes[D]\right)\hat{v}_{1h} + (u_{0h}\otimes d_h).$$

Substitute this into the second equation to get the final spectral equations:

$$A_b\hat{v}_{1h} = \hat{f}_h - \Big(u_{0h}\otimes\left([D]\,d_h\right)\Big) - (u_{1h}\otimes d_h),$$

where

$$A_b = (I_{N-1}\otimes[D]^2) + (B\otimes I_N). \tag{5.8}$$

5.2 Condition number estimates

In this section, we estimate the condition number of the spectral approximations of the various differential operators. The theorem below states that the spectral condition numbers of the discrete spectral operators scale like those of the corresponding elliptic parts. Recall that the result $\kappa(A_h) \leq CN^4$ for the heat equation has already been shown in [49] for the Legendre case and in Section 4.4 for the Chebyshev case.

Theorem 5.1. Let $N \geq 2$. Let A_s be the Chebyshev spectral collocation matrix defined by (5.1) and λ be any eigenvalue of A_s. Then

$$c \leq |\lambda| \leq CN^4.$$

Consequently

$$\kappa(A_s) \leq CN^4.$$

Proof. From (5.1), $\lambda = \gamma + i\mu$, where γ is some eigenvalue of $[D]$ and μ is some eigenvalue of $-[\![D^2]\!]$. Write $\gamma = \gamma_r + i\gamma_i$, where γ_r and γ_i are real. From Lemmas 4.5, 4.16 and 4.10, $\gamma_r \geq c,\ |\gamma| \leq cN^2$ and $c \leq \mu \leq CN^4$ for some positive constants c, C independent of N. Thus

$$|\lambda|^2 = \gamma_r^2 + (\gamma_i + \mu)^2 \leq c^2 + (cN^2 + CN^4)^2 \leq C_1 N^8,$$

and

$$|\lambda|^2 \geq c^2,$$

or equivalently,

$$c \leq |\lambda| \leq CN^4, \qquad \kappa(A_s) \leq CN^4.$$

□

Theorem 5.2. Let $N \geq 2$. Let A_w be the Chebyshev spectral collocation matrix defined

by (5.2) and λ be any eigenvalue of A_w. Then

$$c \leq |\lambda| \leq CN^4.$$

Consequently

$$\kappa(A_w) \leq CN^4.$$

Proof. From (5.2), it follows that

$$\lambda = \gamma^2 + \mu,$$

where $\gamma = \gamma_r + i\gamma_i$ is an eigenvalue of $[D]$ and μ is an eigenvalue of $-[\![D^2]\!]$. A calculation yields

$$|\lambda|^2 = \gamma_r^4 + 2\mu\gamma_r^2 + 2\gamma_r^2\gamma_i^2 + (\mu - \gamma_i^2)^2 \geq \gamma_r^4 + 2\mu\gamma_r^2 \geq c,$$

by Lemmas 4.5 and 4.10. Using Lemmas 4.16 and 4.10, it follows that

$$|\lambda| \leq c^2 + CN^4.$$

Thus

$$c \leq |\lambda| \leq CN^4, \qquad \kappa(A_w) \leq CN^4.$$

□

In following, we present a couple of useful technical results followed by another one which is directly needed for an estimate of the condition number of the Airy spectral operator.

Lemma 5.3. Let $u \in H_w^1(-1,1)$ and $v \in H_{0,w}^1(-1,1)$. Then

$$\left| \int_{-1}^{1} u'(vw)' \right| \leq 2 \|u'\| \, \|v'\|.$$

Proof. The proof follows from a direct calculation:

$$\begin{aligned}
\left|\int_{-1}^{1} u'(vw)'\right| &= \left|\int_{-1}^{1} u'\left(v'w + vw'\right)\right| \\
&\leq \left|\int_{-1}^{1} u'v'w\right| + \left|\int_{-1}^{1} u'vxw^3\right| \\
&\leq \|u'\|\,\|v'\| + \|u'\|\left(\int_{-1}^{1} v^2x^2w^5\right)^{1/2} \\
&\leq \|u'\|\,\|v'\| + \|u'\|\left(\int_{-1}^{1} v^2(1+x^2)w^5\right)^{1/2} \\
&\leq \|u'\|\,\|v'\| + \|u'\|\,\|v'\| \;=\; 2\,\|u'\|\,\|v'\|.
\end{aligned}$$

The last inequality follows from a simple calculation. See (5.61) in [48], for instance. □

Lemma 5.4. Let $v \in H^1_{0,w}(-1,1)$. Then

$$\int_{-1}^{1} \frac{v^2w}{(1-x)^2} \leq \frac{8}{3}\,\|v'\|, \qquad \int_{-1}^{1} \frac{v^2w}{(1-x^2)^2} \leq \frac{2}{3}\,\|v'\|^2.$$

Proof. For $x \in (-1,1)$, it is easy to see that $(1+x)^2 \leq 4$, leading to

$$\frac{1}{\sqrt{1+x}} \leq \frac{4}{(1+x)^{5/2}}.$$

Using the above inequality and a Hardy-type inequality (inequality (13.4) in [7], for instance),

$$\int_{-1}^{1} \frac{v^2}{(1-x^2)^{5/2}} \leq \frac{2}{3}\|v'\|^2,$$

it follows that

$$\int_{-1}^{1} \frac{v^2w}{(1-x)^2} = \int_{-1}^{1} \frac{v^2}{(1-x)^{5/2}\,(1+x)^{1/2}} \leq 4\int_{-1}^{1} \frac{v^2}{(1-x^2)^{5/2}} \leq \frac{8}{3}\,\|v'\|^2.$$

The second inequality of this lemma is exactly the Hardy-type inequality above. □

Proposition 5.5. Let $N \geq 2$ and B be defined in (5.4). Suppose λ is any eigenvalue of

B. Then

$$|\lambda| \le CN^6.$$

Proof. Let v_h be an eigenvector of B corresponding to λ. Let $\zeta \in P_N$ so that $\zeta(\pm 1) = 0$ and $\zeta(x_h) = M^{-1}v_h$, where M is diagonal with entries of the form $1 - x_j$. Define $v(x) = (1-x)\zeta(x) \in P_{N+1}$. Note that $v(x_h) = M\zeta(x_h) = v_h$. Now

$$\lambda v(x_h) = \lambda v_h = Bv_h = \Big(M\,[\![D^3]\!] - 3\,[\![D^2]\!]\Big)\,\zeta(x_h) = v'''(x_h)$$

by (5.3). Observe that $v(\pm 1) = 0 = v'(1)$. By a direct calculation, $v''' = -3\zeta'' + (1-x)\zeta'''$. With these results,

$$\begin{aligned}
\lambda \sum_{j=1}^{N} \frac{|v(x_j)|^2}{(1-x_j)^2}\,\rho_j &= \sum_{j=1}^{N} v'''(x_j)\frac{\overline{v(x_j)}}{(1-x_j)^2}\,\rho_j \\
\lambda \sum_{j=1}^{N} |\zeta(x_j)|^2 \rho_j &= \sum_{j=1}^{N} \Big(-3\zeta''(x_j) + (1-x_j)\zeta'''(x_j)\Big)\frac{\overline{\zeta(x_j)}}{1-x_j}\,\rho_j \\
\lambda \sum_{j=0}^{N} |\zeta(x_j)|^2 \rho_j &= -3\sum_{j=1}^{N} \frac{\zeta''(x_j)\overline{\zeta(x_j)}}{1-x_j}\,\rho_j + \sum_{j=1}^{N} \zeta'''(x_j)\overline{\zeta(x_j)}\rho_j \\
&= -3\sum_{j=0}^{N} \frac{\zeta''(x_j)\overline{\zeta(x_j)}}{1-x_j}\,\rho_j - 3\zeta''(1)\overline{\zeta'(1)}\,\rho_0 + \sum_{j=0}^{N} \zeta'''(x_j)\overline{\zeta(x_j)}\,\rho_j.
\end{aligned}$$

In the first sum of the last equality on the right-hand side, the term $j = 0$ is taken in the sense of a limit since there is division by zero.

Next we estimate the boundary term. Let $\zeta(x) = (1-x)\xi(x)$ with $\xi \in P_{N-1}$. It follows that $\zeta'(1) = -\xi(1)$ and $\zeta''(1) = -2\xi'(1)$. By the trace inequality and Lemma 5.4,

$$|\zeta'(1)|^2 = |\xi(1)|^2 \le c(N-1)\|\xi\|^2 = c(N-1)\int_{-1}^{1} \frac{|\zeta|^2 w}{(1-x)^2} \le CN\|\zeta'\|^2.$$

Similarly, invoking the inverse estimate in addition to the other inequalities,

$$|\zeta''(1)|^2 = 4\,|\xi'(1)|^2 \le c(N-1)\|\xi'\|^2 \le CN^5\|\xi\|^2 \le CN^5\|\zeta'\|^2.$$

The boundary term can now be estimated directly:

$$|\zeta''(1)\zeta'(1)|\,\rho_0 \le CN^{5/2}\|\zeta'\|\,CN^{1/2}\|\zeta'\|\,\frac{\pi}{2N} = cN^2\|\zeta'\|^2 \le CN^6\|\zeta\|^2.$$

Finally, the magnitude of the eigenvalue can be estimated using Lemmas 5.3 and 5.4 and the fact that integration can be evaluated exactly for any integrand of degree $2N-1$ or lower:

$$\begin{aligned}
|\lambda|\sum_{j=0}^{N}\zeta^2(x_j)\rho_j &\le 3\left|\int_{-1}^{1}\frac{\zeta''\bar{\zeta}w}{1-x}\right| + \left|\int_{-1}^{1}\zeta'''\bar{\zeta}w\right| + CN^6\|\zeta\|^2 \\
&\le 3\,\|\zeta''\|\left(\int_{-1}^{1}\frac{\bar{\zeta}^2 w}{(1-x)^2}\right)^{1/2} + \left|\int_{-1}^{1}\zeta''(\bar{\zeta}w)'\right| + CN^6\|\zeta\|^2 \\
&\le cN^2\|\zeta'\|\,\|\zeta'\| + c\,\|\zeta''\|\|\bar{\zeta}'\| + CN^6\|\zeta\|^2 \\
&\le CN^6\|\zeta\|^2.
\end{aligned}$$

In the last two lines, the inverse estimate has been invoked several times. It follows from the equivalence of the discrete and weighed L^2 norms that $|\lambda| \le CN^6$.

□

We remark that It should be possible to show lower bounds using Proposition 4.15 and Lemma 4.16.

Theorem 5.6. Let $N \ge 2$. Let A_a be the Chebyshev spectral collocation matrix defined by (5.5) and λ be any eigenvalue of A_a. Then

$$|\lambda| \le CN^6.$$

Proof. From (5.5), it follows that

$$\lambda = \gamma^2 + \mu,$$

where $\gamma = \gamma_r + i\gamma_i$ is an eigenvalue of $[D]$ and μ is an eigenvalue of B defined in (5.4).

Using Lemma 4.16 and Proposition 5.5, it follows that

$$|\lambda| \leq CN^6.$$

□

Next, we present a lemma which is needed for an estimate of the condition number of the beam spectral operator.

Lemma 5.7. Let $N \geq 2$ and B be defined in (5.7). Suppose λ is any eigenvalue of B. Then

$$\lambda \leq CN^8.$$

Proof. Let v_h be an eigenvector of B corresponding to λ. Let $\zeta \in P_N$ so that $\zeta(\pm 1) = 0$ and $\zeta(x_h) = M^{-1}v_h$, where M is diagonal with diagonal entries of the form $1 - x_j^2$. Define $v(x) = (1-x^2)\zeta(x) \in P_{N+2}$. Note that $v(x_h) = M\zeta(x_h) = v_h$. Now

$$\lambda v(x_h) = \lambda v_h = Bv_h = \left(M[\![D^4]\!] - 8X[\![D^3]\!] - 12[\![D^2]\!]\right)\zeta(x_h) = v''''(x_h)$$

by (5.6). Observe that $v(\pm 1) = 0 = v'(\pm 1)$. By a direct calculation, $v'''' = -12\zeta'' - 8x\zeta''' + (1-x^2)\zeta''''$. With these results,

$$\begin{aligned}
\lambda \sum_{j=1}^{N-1} \frac{|v(x_j)|^2}{(1-x_j^2)^2}\rho_j &= \sum_{j=1}^{N-1} \frac{v''''(x_j)\overline{v(x_j)}}{(1-x_j^2)^2}\rho_j \\
\lambda \sum_{j=1}^{N-1} |\zeta(x_j)|^2\rho_j &= -12\sum_{j=1}^{N-1} \frac{\zeta''(x_j)\overline{\zeta(x_j)}}{1-x_j^2}\rho_j - 8\sum_{j=1}^{N-1} \frac{x_j\zeta'''(x_j)\overline{\zeta(x_j)}}{1-x_j^2}\rho_j + \sum_{j=1}^{N-1} \zeta''''(x_j)\overline{\zeta(x_j)}\,\rho_j \\
\lambda \sum_{j=0}^{N} |\zeta(x_j)|^2\rho_j &= -12\sum_{j=0}^{N} \frac{\zeta''(x_j)\overline{\zeta(x_j)}}{1-x_j^2}\rho_j - 8\sum_{j=0}^{N} \frac{x_j\zeta'''(x_j)\overline{\zeta(x_j)}}{1-x_j^2}\rho_j + \sum_{j=0}^{N} \zeta''''(x_j)\overline{\zeta(x_j)}\,\rho_j \\
&\quad + \frac{3\pi}{N}\left(\zeta''(-1)\overline{\zeta'(-1)} - \zeta''(1)\overline{\zeta'(1)}\right) + \frac{2\pi}{N}\left(\zeta'''(-1)\overline{\zeta'(-1)} - \zeta'''(1)\overline{\zeta'(1)}\right).
\end{aligned}$$

Let $\zeta(x) = (1-x^2)\xi(x)$ with $\xi \in P_{N-2}$. Using a similar technique as before, we estimate

the boundary terms:

$$|\zeta'(\pm 1)|^2 = 4\,|\xi(\pm 1)|^2 \le CN\|\zeta'\|^2, \qquad |\zeta''(\pm 1)|^2 = |2\xi(\pm 1) + 4(\pm 1)\xi'(\pm 1)|^2 \le CN^5\|\zeta'\|^2,$$

and

$$|\zeta'''(\pm 1)|^2 = |6\xi'(\pm 1) + 6(\pm 1)\xi''(\pm 1)|^2 \le CN^9\|\zeta'\|^2,$$

leading to a final upper bound of all boundary terms of $CN^8\|\zeta\|^2$.

For the final estimate of the eigenvalue, again use Lemmas 5.3 and 5.4 and the fact that the integration can be evaluated exactly by summation since the integrand is of degree at most $2N-1$ to get

$$\begin{aligned}
|\lambda| \sum_{j=0}^{N} |\zeta(x_j)|^2 \rho_j &\le 12 \left| \int_{-1}^{1} \frac{\zeta'' \bar{\zeta} w}{1-x^2} \right| + 8 \left| \int_{-1}^{1} \frac{x\zeta''' \bar{\zeta} w}{1-x^2} \right| + \left| \int_{-1}^{1} \zeta'''' \bar{\zeta} w \right| + CN^8\|\zeta\|^2 \\
&\le 12\,\|\zeta''\| \left(\int_{-1}^{1} \frac{|\zeta|^2 w}{(1-x^2)^2} \right)^{1/2} + 8\,\|\zeta'''\| \left(\int_{-1}^{1} \frac{|\zeta|^2 w}{(1-x^2)^2} \right)^{1/2} \\
&\qquad + \left| \int_{-1}^{1} \zeta'''(\bar{\zeta} w)' \right| + CN^8\|\zeta\|^2 \\
&\le CN^2\|\zeta'\|^2 + CN^4\|\zeta'\|^2 + \|\zeta'''\|\,\|\zeta'\| + CN^8\|\zeta\|^2 \\
&\le CN^8\|\zeta\|^2.
\end{aligned}$$

By the equivalence of the discrete and weighed L^2 norms, $|\lambda| \le CN^8$.

□

We remark that It should be possible to show lower bounds using Proposition 4.15 and Lemma 4.16.

Theorem 5.8. Let $N \ge 2$. Let A_b be the Chebyshev spectral collocation matrix defined by (5.8) and λ be any eigenvalue of A_b. Then

$$\lambda \le CN^8.$$

Proof. From (5.8), it follows that

$$\lambda = \gamma^2 + \mu,$$

where $\gamma = \gamma_r + i\gamma_i$ is an eigenvalue of $[D]$ and μ is an eigenvalue of B defined in (5.7). Using Lemma 4.16 and Proposition5.7, it follows that

$$|\lambda| \leq CN^8.$$

□

5.3 Spectral Convergence

In this section, we discuss space-time spectral convergence of our method for the Schrodinger and wave equations.

Theorem 5.9. Let u be the solution of the Schrodinger equation. Assume u is separately analytic in each variable. Let $N \geq 2$ and $\hat{u}_h$ be the solution of (5.1). Define the error vector E_h as the difference of u evaluated at the collocation points and $\hat{u}_h$. Then

$$|W^{1/2} E_h|_2 \leq cN^{3.5} e^{-CN}.$$

The proof of spectral convergence for the Schrodinger equation is almost identical to that of the heat equation in Section 4.4 and is omitted. What is perhaps surprising is that the method of proof is so similar despite the fact that this PDE is dispersive and has completely different properties from those of the heat equation which is diffusive.

Theorem 5.10. Let u be the solution of the wave equation. Assume u is separately analytic in each variable. Let $N \geq 2$ and $\hat{v}_{1h}$ be the solution of the space-time method with matrix defined by (5.2). Define the error vector E_h as the difference of u evaluated

at the collocation points and $\hat{v}_{1h}$. Then

$$|W^{1/2}E_h|_2 \leq cN^{4.5}e^{-CN}.$$

Proof. Define

$$u_h(t) = \begin{bmatrix} u(x_1,t) \\ \vdots \\ u(x_{N-1},t) \end{bmatrix}, \qquad f_h(t) = \begin{bmatrix} f(x_1,t) \\ \vdots \\ f(x_{N-1},t) \end{bmatrix}.$$

A semi-discrete approximation of the wave equation is

$$u_h''(t) = \sum_{j=0}^{N} \Big(Au_h(t_j) + f_h(t_j) \Big) \ell_j(t), \qquad u_h(-1) = u_{0h},\ u_h'(-1) = u_{1h},$$

where $A = [\![D^2]\!]$. Hence

$$u_h''(t_k) = Au_h(t_k) + f_h(t_k), \qquad 0 \leq k \leq N-1.$$

Define the error function $e_h(t) = u_h(t) - u(x_h,t)$ with components $e_j(t) = (e_h(t))_j$. Using the above equation, it is easy to see that the error satisfies, for $0 \leq k \leq N-1$,

$$e_h''(t_k) = Ae_h(t_k) + r(t_k), \qquad r(t_k) = Au(x_h,t_k) - u_{xx}(x_h,t_k). \tag{5.9}$$

For any analytic z such that $z(-1) = 0$, recall the definition of the interpolant

$$\mathcal{I}_N z(t) = \sum_{j=0}^{N-1} z(t_j)\ell_j(t).$$

For $0 \leq k \leq N-1$,

$$z'(t_k) = (\mathcal{I}_N z)'(t_k) + \tilde{\epsilon}_k = \Big([D](\mathcal{I}_N z)(t_h) \Big)_k + \tilde{\epsilon}_k = \Big([D] z(t_h) \Big)_k + \tilde{\epsilon}_k, \tag{5.10}$$

where $\tilde{\epsilon}_k = (z - \mathcal{I}_N z)'(t_k)$ satisfies

$$|\tilde{\epsilon}_k| \leq cN^2 e^{-CN}$$

according to [61]. Take $z(t) = e_j(t)$ in (5.10), observing that $e_j(-1) = 0$, then

$$e_j'(t_k) = \Big([D]e_j(t_h)\Big)_k + \epsilon_{1jk}, \tag{5.11}$$

where $|\epsilon_{1jk}| \leq cN^2 e^{-CN}$. Next take $z(t) = e_j'(t)$ in (5.10), noting that $e_j'(-1) = 0$, then

$$e_j''(t_k) = \Big([D]e_j'(t_h)\Big)_k + \epsilon_{2jk}, \tag{5.12}$$

where $|\epsilon_{2jk}| \leq cN^2 e^{-CN}$. Considering (5.9) together with (5.11) and (5.12), we have

$$\Big([D]e_j(t_h)\Big)_k + \epsilon_{1jk} = (e_h'(t_k))_j, \tag{5.13}$$

$$\Big([D]e_j'(t_h)\Big)_k + \epsilon_{2jk} = (Ae_h(t_k))_j + r_j(t_k), \tag{5.14}$$

where residual vectors $r_j(t_h) = Au(x_j, t_h) - u_{xx}(x_j, t_h)$. Define the long vector

$$\tilde{R}_h = \begin{bmatrix} r_1(t_h) \\ \vdots \\ r_{N-1}(t_h) \end{bmatrix},$$

and

$$E_h = \begin{bmatrix} e_1(t_h) \\ \vdots \\ e_{N-1}(t_h) \end{bmatrix},$$

in vector notation, the equations (5.13) and (5.14) are

$$(I_{N-1} \otimes [D])E_h = E'_h - \epsilon_1, \qquad (I_{N_1} \otimes [D])E'_h = (A \otimes I_N)E_h + \tilde{R}_h - \epsilon_2,$$

where ϵ_1, ϵ_2 are long vectors formed by stacking together vectors $[e_{pj0}, \ldots, \epsilon_{pj,N-1}]^T$ for $p = 1, 2$ and $1 \leq j \leq N-1$; and each component of E'_h has the form $e'_j(t_k)$. Combine these two equations to obtain

$$A_w E_h = R_h := \tilde{R}_h - \epsilon_2 - (I_{N-1} \otimes [D])\epsilon_1.$$

Using Lemma 4.16 and the above estimates, it follows that $|R|_\infty \leq cN^4 e^{-CN}$. Apply the result of Theorem 5.10 and proceed as in Section 4.4 to get the desired error estimate. □

5.4 Nonlinear PDEs

The purpose of this section is to show that it is simple, in a few lines of code in the spirit of [72], to adapt the above methodology to solve some of the most common nonlinear PDEs with spectral space-time convergence. In [49] and in Section 4.4 we had considered the Allen-Cahn equation and Burgers' equation. We now look at some other nonlinear PDEs.

5.4.1 Nonlinear reaction diffusion equation

Consider

$$u_t = u_{xx} + \lambda e^u + f(x,t), \quad \text{on } (-1,1)^2,$$

with initial condition $u(x,-1) = u_0(x)$ and homogeneous Dirichlet boundary conditions. Here λ is a positive constant. The spectral scheme is

$$(I_{N+1} \otimes D)\, u_h = (D^2 \otimes I_{N+1})\, u_h + \lambda e^{u_h} + f_h,$$

where f_h is f evaluated at collocation points. Deleting the known boundary and initial values, the final scheme reads

$$[(I_{N-1} \otimes [D]) - ([\![D^2]\!] \otimes I_N)]\, \hat{u}_h - \lambda e^{\hat{u}_h} = \hat{f}_h - (u_{0h} \otimes d_h),$$

where u_{0h} is u_0 evaluated at the interior collocation points. This nonlinear system can be solved using the simple iteration ($k \geq 0$)

$$[(I_{N-1} \otimes [D]) - ([\![D^2]\!] \otimes I_N)]\, \hat{u}_h^{(k+1)} = \lambda e^{\hat{u}_h^{(k)}} + \hat{f}_h - \Big(u_{0h} \otimes d_h\Big).$$

5.4.2 Nonlinear Schrodinger equation

Consider

$$iu_t = -u_{xx} + |u|^2 u + f(x,t), \quad \text{on } (-1,1)^2,$$

with initial condition $u(x,-1) = u_0(x)$ and homogeneous Dirichlet boundary conditions. The spectral scheme is

$$i(I_{N+1} \otimes D)\, u_h = -(D^2 \otimes I_{N+1}) u_h + |u_h|^2 u_h + f_h,$$

where f_h is f evaluated at collocation points. Deleting the known boundary and initial values, the final scheme reads

$$i[(I_{N-1} \otimes [D]) + ([\![D^2]\!] \otimes I_N)]\, \hat{u}_h - |\hat{u}_h|^2 \hat{u}_h = \hat{f}_h - (u_{0h} \otimes d_h),$$

where u_{0h} is u_0 evaluated at the interior collocation points. This nonlinear system can be solved using the simple iteration ($k \geq 0$) with relaxation:

$$i[(I_{N-1} \otimes [D]) + ([\![D^2]\!] \otimes I_N)]\, \tilde{u}_h^{(k+1)} - |\hat{u}_h^{(k)}|^2 \tilde{u}_h^{(k+1)} + \hat{f}_h - \Big(u_{0h} \otimes d_h\Big), \qquad \hat{u}_h^{(k+1)} = \frac{\tilde{u}_h^{(k+1)} + u_h^{(k)}}{2}.$$

5.4.3 Sine–Gordon equation

The Sine–Gordon equation is

$$u_{tt} = u_{xx} + \sin u + F(x,t), \quad \text{on } (-1,1)^2,$$

with initial conditions $u(x,-1) = u_0(x)$ and $u_t(x,-1) = u_1(x)$ and homogeneous Dirichlet boundary conditions. The spectral scheme is

$$(I_{N+1} \otimes D^2)\, u_h = (D^2 \otimes I_{N+1})\, u_h + \sin u_h + f_h.$$

Deleting the known boundary and initial values, the final scheme reads

$$[(I_{N-1} \otimes [D]^2) - (\llbracket D^2 \rrbracket \otimes I_N)]\, \hat{u}_h - \sin \hat{u}_h = \hat{f}_h - \Big(u_{0h} \otimes ([D]\, d_h)\Big) - (u_{1h} \otimes d_h).$$

This nonlinear system can be solved using the iteration ($k \geq 0$)

$$[(I_{N-1} \otimes [D]^2) - (\llbracket D^2 \rrbracket \otimes I_N)]\hat{u}_h^{(k+1)} = \sin \hat{u}_h^{(k)} + \hat{f}_h - \Big(u_{0h} \otimes ([D]\, d_h)\Big) - (u_{1h} \otimes d_h).$$

5.4.4 KdV equation

The KdV equation is

$$u_t + uu_x + u_{xxx} = F(x,t), \quad \text{on } (-1,1)^2,$$

with initial condition $u(x,-1) = u_0(x)$ and boundary conditions $u(-1,t) = 0 = u(1,t) = u_x(1,t)$. The spectral scheme is

$$(I_{N+1} \otimes D)u_h + \mathrm{diag}(u_h)(D \otimes I_{N+1})u_h + (D^3 \otimes I_{N+1})u_h = f_h.$$

Let B be the spectral third derivative (5.4) defined for the Airy operator. The final system, removing the known boundary and initial values, becomes

$$[(I_{N-1} \otimes [D]) + (B \otimes I_N)]\hat{u}_h + \mathrm{diag}((\llbracket D \rrbracket \otimes I_N)\hat{u}_h)\hat{u}_h = \hat{f}_h - (u_{0h} \otimes d_h).$$

This can be solved using the iteration ($k \geq 0$)

$$[(I_{N-1} \otimes [D]) + (B \otimes I_N)]\hat{u}_h^{(k+1)} + \mathrm{diag}((\llbracket D \rrbracket \otimes I_N)\hat{u}_h^{(k)})\hat{u}_h^{(k+1)} = \hat{f}_h - (u_{0h} \otimes d_h).$$

5.4.5 Kuramoto–Sivashinsky equation

The Kuramoto–Sivashinsky equation reads

$$u_t + u_{xxxx} + u_{xx} + uu_x = F(x,t), \quad \text{on } (-1,1)^2,$$

with initial condition $u(x,-1) = u_0(x)$ and homogeneous Dirichlet boundary conditions. The scheme is then

$$\Big((I_{N-1} \otimes [D]) + (B + \llbracket D^2 \rrbracket) \otimes I_N)\Big)\ \hat{u}_h + \frac{1}{2}(\llbracket D \rrbracket \otimes I_N)\ \hat{u}_h^2 = \hat{f}_h - (u_{0h} \otimes d_h).$$

This nonlinear system can be solved using the iteration ($k \geq 0$)

$$\Big((I_{N-1} \otimes [D]) + (B + \llbracket D^2 \rrbracket) \otimes I_N)\Big)\ \hat{u}_h^{(k+1)} + \mathrm{diag}\Big((\llbracket D \rrbracket \otimes I_N)\ \hat{u}_h^{(k)}\Big)\hat{u}_h^{(k+1)} = \hat{f}_h - (u_{0h} \otimes d_h).$$

5.4.6 Cahn-Hilliard equation

The Cahn-Hilliard equation is

$$u_t - (-u_{xx} + u^3 - u)_{xx} = F(x,t)$$

with initial condition $u(x,-1) = u_0(x)$ and boundary conditions

$$u_x(\pm 1, t) = 0 = u_{xxx}(\pm 1, t).$$

The full scheme, using Legendre space-time collocation, is

$$\Big((I_{N+1} \otimes D) + \Big((D^4 + D^2) \otimes I_{N+1}\Big)\Big)\, u_h - (D^2 \otimes I_N)\, u_h^3 = f_h.$$

Let $B = W^{-1} D^T W D$, where W is the diagonal matrix whose diagonal entries are the weights of the collocation scheme. It is known ([48]) that $-B$ is a spectral approximation of the second derivative for functions whose derivative vanishes at the boundary. The spectral equations for $\hat{u}_h$, the entries of u_h removing the initial values, become

$$\Big((I_{N+1} \otimes [D]) + (B^2 - B) \otimes I_N)\Big)\, \hat{u}_h + (B \otimes I_N)\, \hat{u}_h^3 = \hat{f}_h - (u_{0h} \otimes d_h).$$

This nonlinear system can be solved iteratively. Let $\tilde{D}$ be D except that the first and last rows are replaced by a row of zeroes. This is a spectral approximation of the first derivative for functions whose derivative vanish at the boundary. There are several ways to discretize $(u^3)_{xx} = 2uu_x^2 + u^2 u_{xx}$. We attempted two, one of which worked, but not the other. The simple scheme

$$\Big((I_{N+1} \otimes [D]) + (B^2 - B) \otimes I_N)\Big)\, \hat{u}_h^{(k+1)} - 6\,\mathrm{diag}\Big((\tilde{D} \otimes I_N)\hat{u}_h^{(k)}\Big)^2 \hat{u}_h^{(k+1)}$$
$$+3\,\mathrm{diag}(\hat{u}_h^{(k)})^2 (B \otimes I_N)\hat{u}_h^{(k+1)} = \hat{f}_h - (u_{0h} \otimes d_h),$$

did not seem to converge. The following iteration with relaxation does seem to work ($k \geq 0$):

$$\Big((I_{N+1} \otimes [D]) + (B^2 - B) \otimes I_N)\Big)\ \tilde{u}_h^{(k+1)} - 6\,\mathrm{diag}(\hat{u}_h^{(k)})\mathrm{diag}\Big((\tilde{D} \otimes I_N)\hat{u}_h^{(k)}\Big)(\tilde{D} \otimes I_N)\tilde{u}_h^{(k+1)}$$
$$+3\,\mathrm{diag}(\hat{u}_h^{(k)})^2(B \otimes I_N)\tilde{u}_h^{(k+1)} = \hat{f}_h - (u_{0h} \otimes d_h),$$
$$\hat{u}_h^{(k+1)} = \frac{\tilde{u}_h^{(k+1)} + u_h^{(k)}}{2},$$

where u_{0h} is the initial data evaluated at all spatial collocation points. It is beyond the scope of this thesis to discuss convergence theories of the schemes in this section.

5.5 Numerical Results

We implemented a very simple space-time Legendre and Chebyshev collocation method for each PDE discussed in this paper in MATLAB. Results for the Chebyshev case are reported below. Almost identical results hold for the Legendre case and they are not given.

The convergence for the Schrodinger equation

$$u_t = iu_{xx} + f,$$

with boundary conditions $u(\pm 1, t) = 0$ and initial condition $u(x, -1) = u_0(x)$. Take f so that the exact solution is $u(x, t) = e^{x+t}\sin(\pi t/2)\sin \pi x$. Spectral convergence is clearly illustrated in the left figure of Figure 5.1. The error is the largest error of the numerical solution at the Chebyshev nodes at the final time $t = 1$. Note that the error is $O(10^{-14})$ at $N = 18$ which corresponds to a system with 306 unknowns.

The convergence of the Chebyshev collocation method for the wave equation

$$u_{tt} = u_{xx} + f,$$

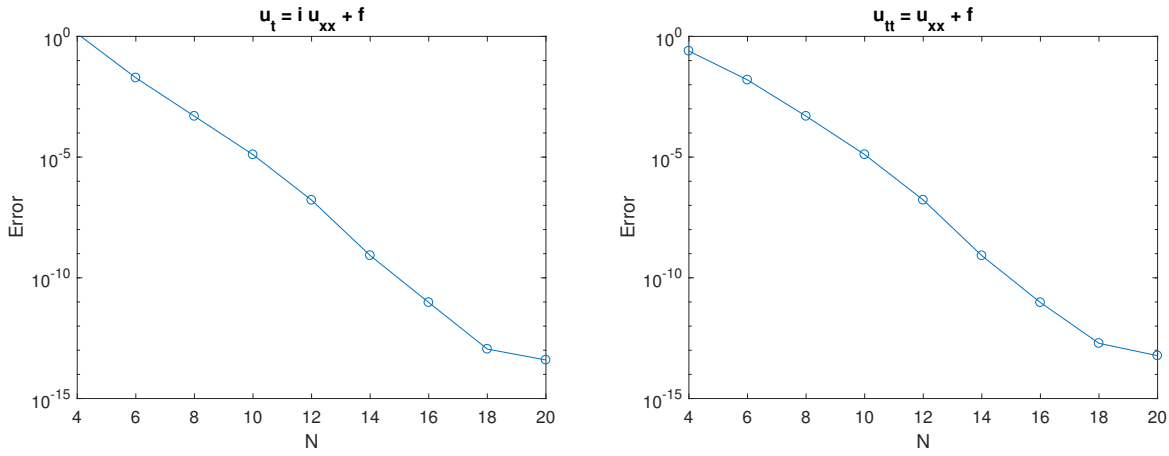

Figure 5.1: Convergence of Chebyshev collocation method for the Schrodinger (left) and wave (right) equations.

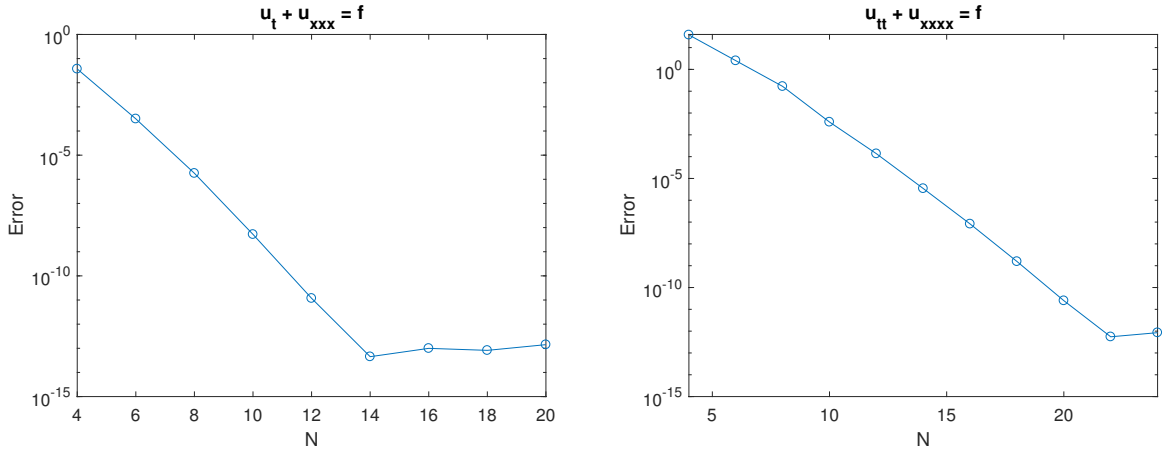

Figure 5.2: Convergence of Chebyshev collocation method for the Airy (left) and beam (right) equations.

with boundary conditions $u(\pm 1, t) = 0$ and initial conditions $u(x, -1) = u_0(x)$ and $u_t(x, -1) = u_1(x)$ can be found in the right figure of Figure 5.1. Here we take f so that the exact solution is the same as above.

For the Airy equation

$$u_t + u_{xxx} = f,$$

with boundary conditions $u(\pm 1, t) = 0 = u_x(1, t)$ and initial condition $u(x, 0) = u_0(x)$, with the same exact solution as before, spectral convergence of the space-time Chebyshev collocation method is shown in the left figure of Figure 5.2.

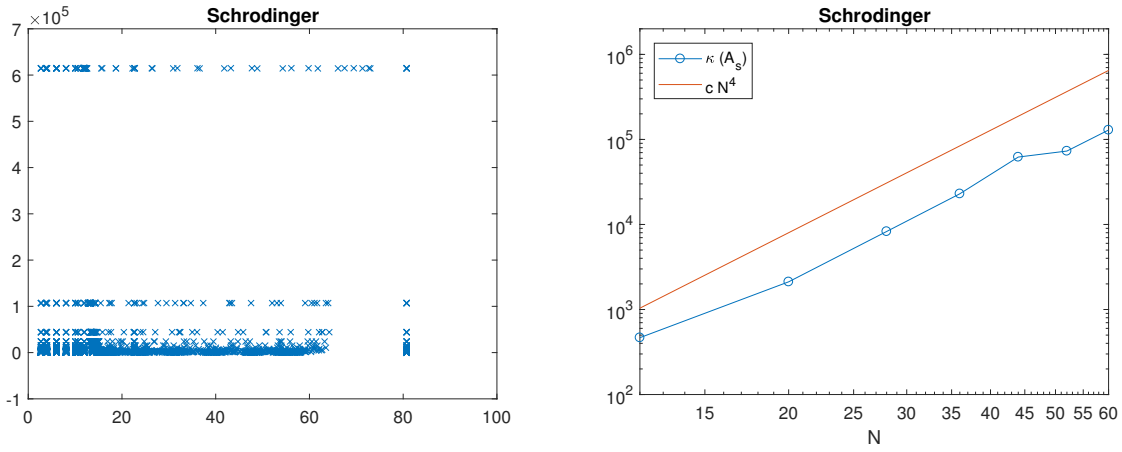

Figure 5.3: Spectrum (left) and spectral condition number (right) for the Schrodinger operator A_s.

Next, consider the beam equation

$$u_{tt} + u_{xxxx} = f,$$

with clamped boundary conditions $u(\pm 1, t) = 0 = u_x(\pm 1, t)$ and initial conditions $u(x, -1) = u_0(x),\ u_t(x, -1) = u_1(x)$. Take f so that the exact solution is $e^{x+t}\sin(\pi t/2)\sin^2(\pi x)$, the spectral convergence of the Chebyshev collocation method can be seen in the right figure of Figure 5.2.

The spectrum of the various spectral Chebyshev operators for the case $N = 60$ and plots of the spectral condition numbers as functions of N are shown in Figures 5.3, 5.4, 5.5, 5.6.

Now we move onto nonlinear PDEs. For all nonlinear PDEs, we take as initial guess the zero function and use the iteration defined for each nonlinear PDE. The iteration is stopped whenever the infinity norm of the difference of two consecutive iterates are smaller than $\epsilon = 10^{-13}$. Consider first the nonlinear reaction diffusion equation

$$u_t = u_{xx} + \lambda e^u + f,$$

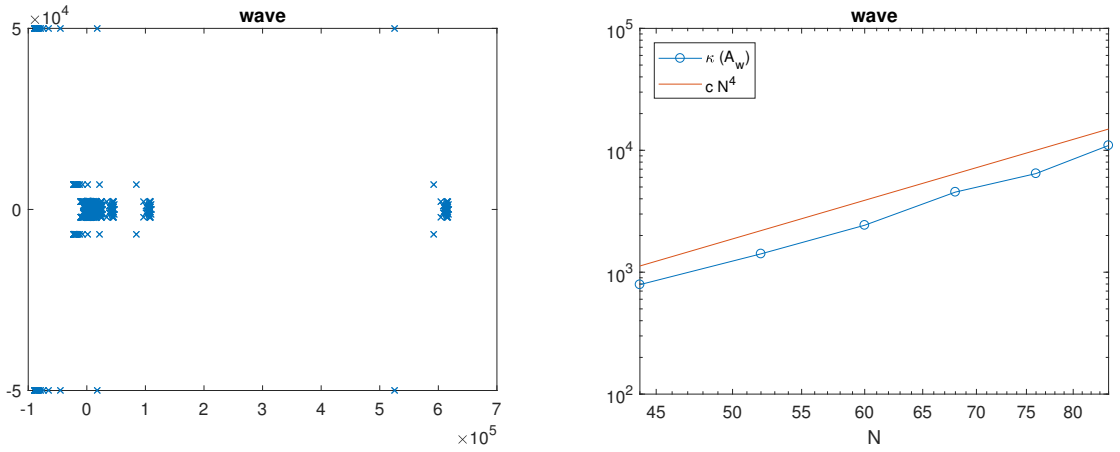

Figure 5.4: Spectrum (left) and spectral condition number (right) for the wave operator A_w.

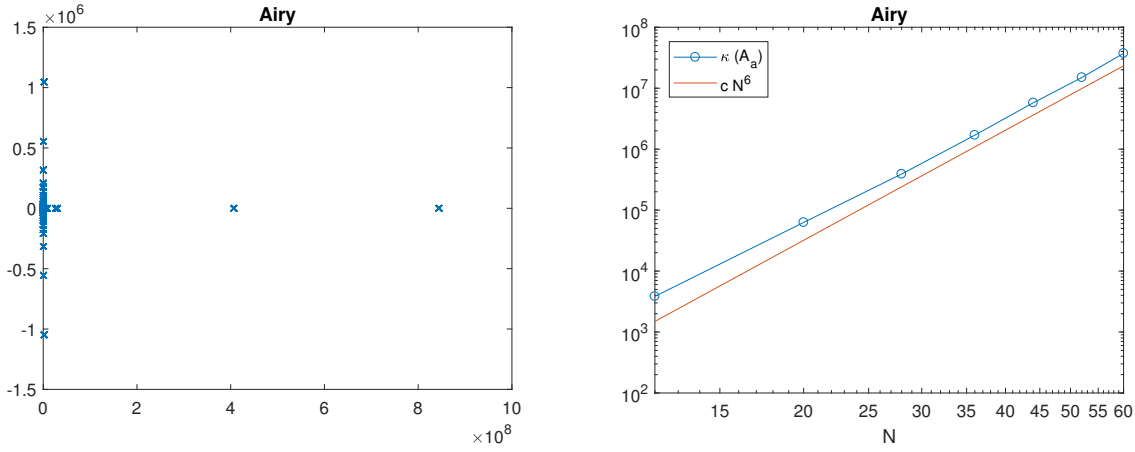

Figure 5.5: Spectrum (left) and spectral condition number (right) for the Airy operator A_a.

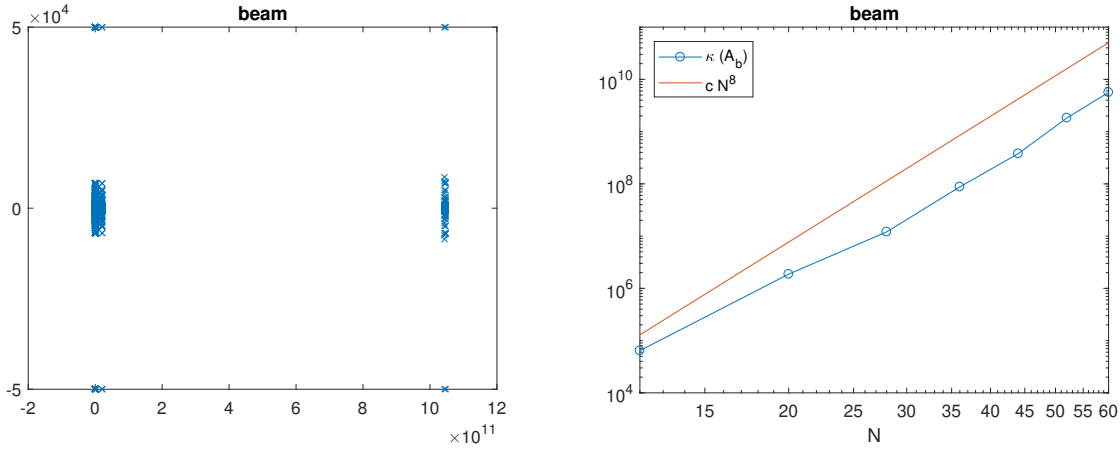

Figure 5.6: Spectrum (left) and spectral condition number (right) for the beam operator A_b.

with homogeneous boundary conditions. Take $\lambda = 0.5$ and f so that the exact solution is $u(x,t) = e^{x+t}\cos(\pi x/2)$. See the left figure of Figure 5.7 for the convergence.

Next consider the nonlinear Schrodinger equation

$$iu_t - u_{xx} + |u|^2 u = f,$$

with homogeneous Dirichlet boundary conditions. Take f so that the exact solution is $u(x,t) = e^{x+t}\sin(\pi x)$. See the right figure of Figure 5.7 for the convergence.

Next consider the Sine–Gordon equation

$$u_{tt} = u_{xx} + \sin u + f,$$

with homogeneous Dirichlet boundary conditions. Take f so that the exact solution is $u(x,t) = e^{x+t}\sin \pi x$. See the right figure of Figure 5.8 for the convergence.

Consider first the KdV equation

$$u_t + uu_x + u_{xxx} = f,$$

with boundary conditions $u(\pm 1, t) = 0 = u_x(1,t)$. Take f so that the exact solution is $u(x,t) = \cos(x-t)\ (x-1)^2(x+1)$. See the left figure of Figure 5.8 for the convergence.

Next, consider the nonlinear Kuramoto-Sivashinsky equation

$$u_t + u_{xxxx} + u_{xx} + uu_x = f,$$

with clamped boundary conditions $u(\pm 1, t) = 0 = u_x(\pm 1, t)$ and initial conditions $u(x,-1) = u_0(x)$. Take f so that the exact solution is $e^{x+t}\sin^2(\pi x)$, the spectral convergence of the Chebyshev collocation method can be seen in the left figure of Figure 5.9.

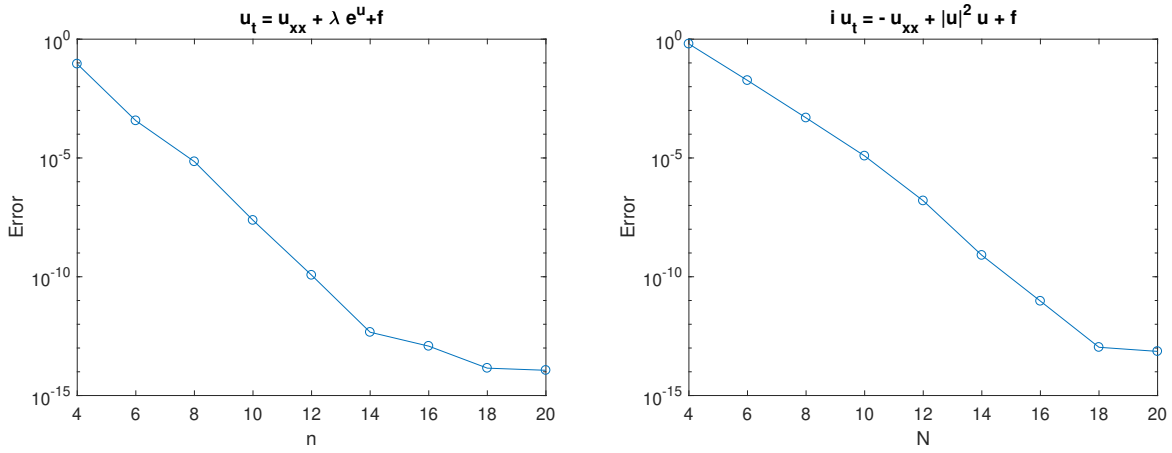

Figure 5.7: Convergence of Chebyshev collocation method for the nonlinear reaction diffusion equation (left) and nonlinear Schrodinger equation (right).

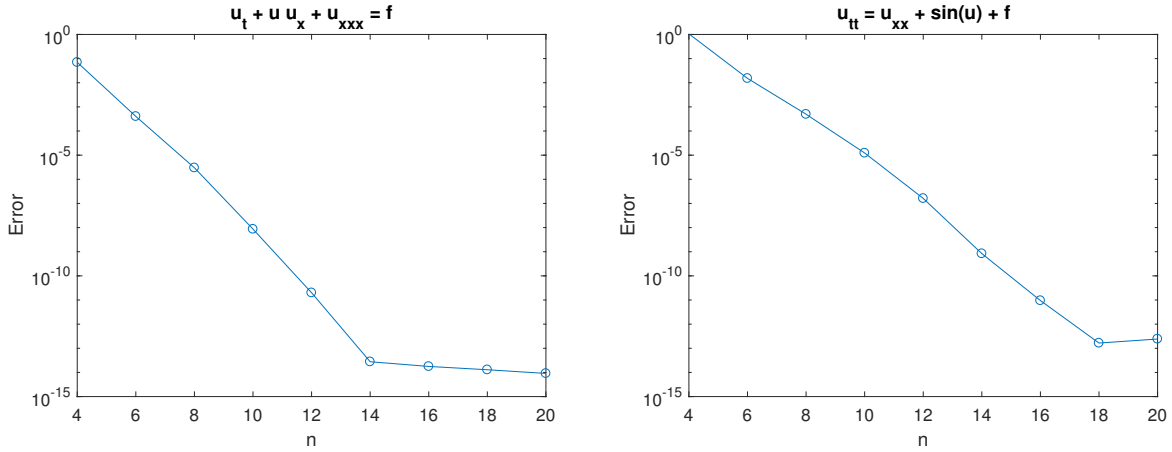

Figure 5.8: Convergence of Chebyshev collocation method for the KdV (left) and Sine–Gordon (right) equations.

Finally, consider the Cahn-Hilliard equation

$$u_t + u_{xxxx} + u_{xx} - (u^2)_{xx} = f,$$

with boundary conditions $u_x(\pm 1, t) = 0 = u_{xxx}(\pm 1, t)$ and initial conditions $u(x, -1) = u_0(x)$. Take f so that the exact solution is $\cos(t)\cos(\pi x)$, the spectral convergence of the Chebyshev collocation method can be seen in the right figure of Figure 5.9. This PDE was the most difficult to solve. The stopping criterion was reduced to $\epsilon = 10^{-9}$.

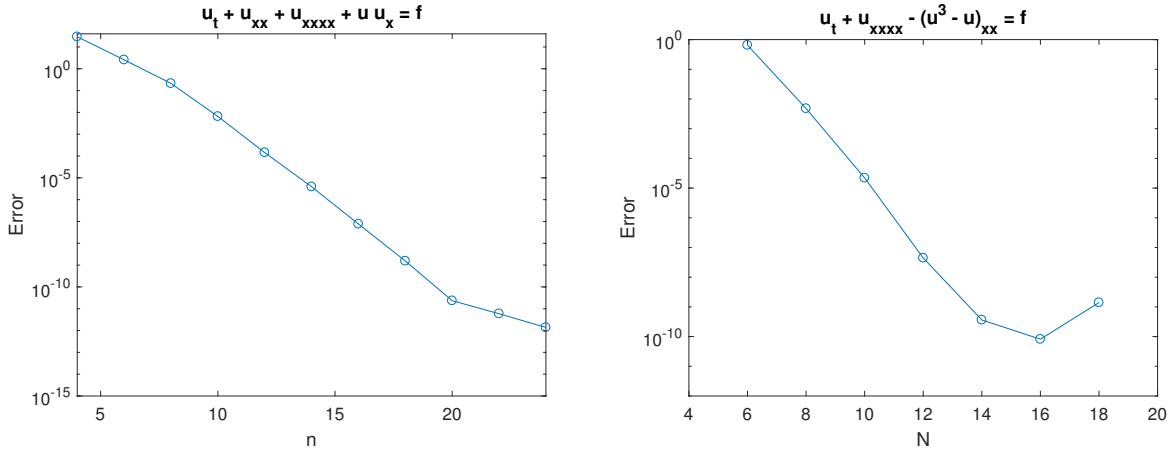

Figure 5.9: Convergence of Chebyshev collocation method for the Kuramoto-Sivashinsky (left) and Cahn–Hilliard (right) equations.

6

Conclusion and future work

In Chapter 2 we have given a local superlinear convergence theory for the solution of a system of nonlinear equations by the basic Broyden's method and the minimizer of a nonlinear function by the BFGS method using Kantorovich-type assumptions, i.e., where all assumptions are about the initial iterate and its neighbourhood. The main point is that the assumptions can be verified in practice. Also, our theories are simple in the sense that they contain as few constants as possible.

In a continuation of Chapter 2, in Chapter 3 we have given a local superlinear convergence theory for the solution of the problem of finding a local minimum of a nonlinear function by using Kantorovich-type assumptions. The symmetric scaled Perry NCG method and generalized scaled memoryless BFGS method are considered. Our theories are simple in the sense that they contain as few constants as possible.

There are many other directions for further research in this area. For instance, the Jacobian matrix for a nonlinear system or the Hessian in the case of unconstrained minimization may be sparse or may have a special structure. [77] has a convergence theory for quasi-Newton methods which maintain the sparsity or special structure. A similar result using Kantorovich-type assumptions would be desirable. Another possible future work is to relax the condition that the Jacobian matrix about the initial point is non-singular, or

the condition that the Hessian of the objective function is positive definite. See [20] for some early work in this direction. Next, two convergence theories for functions, which are not smooth can be found in [60] and [44]. It would be desirable to extend these results for the case of Kantorovich-type assumptions. Smale gives an amazing convergence of Newton's iteration where all assumptions are at the initial iterate - no assumption is necessary in a neighbourhood about the initial iterate. See Chapter 8 in [8]. This theory has been extended to a secant method in [76]. It appears to be an open problem whether this theory carries over to the methods of Broyden and BFGS. Also we have only considered the Perry nonlinear conjugate gradient method. There are many other classes of nonlinear conjugate gradient methods that can be examined. They may require a different technique of proof if they are not of the quasi-Newton type.

In Chapter 4, we have shown that the space-time Chebyshev collocation method of Tang and Xu [71] converges spectrally in both space and time for the heat equation. The condition number of this method is shown to be bounded by $O(N^4)$. We have also proposed another space-time spectral collocation which is easier to implement and has similar characteristics as the first one. Some simple numerical experiments verify the theoretical results. Numerical results for the viscous Burgers and Allen–Cahn equations demonstrate the potential of this method for nonlinear PDEs.

In Chapter 5, we have also extended our analysis for other standard linear PDEs (Schrodinger, Airy, wave and beam equations) and conducted numerical experiments for common nonlinear PDEs (nonlinear diffusion, KdV, Sine–Gordon, Kuramoto–Shivashinsky and Cahn–Hilliard) with similar results. It is remarkable that space-time spectral methods work so well for these different classical PDEs with different features: diffusion, dispersion, nonlinear advection, etc.

Although we have only considered one spatial dimension, the method generalizes to the spatial domain $(-1,1)^d$ immediately for $d \geq 1$. Also, the implementation of the collocation method for general linear variable coefficient PDEs with standard linear boundary

conditions is quite straightforward. The most glaring problem of our method is that the unknowns at all intermediate times are solved simultaneously. For a PDE in d space dimensions, the method requires the numerical solution of a discrete problem with N^{d+1} unknowns, where N is the number of unknowns in each dimension. While direct solvers based on 1D matrix factorization (Bartel-Stewart or other similar algorithms) work well for 1D and 2D problems, it is desirable to come up with more efficient solvers for 3D problems. One potential method is to use, for instance, the backward Euler method to solve the discrete heat equation at the collocation points in time. This gives a good initial guess to the solution of either (4.11) or (4.15). Then an iterative method can be used to improve the accuracy. The original work [71] used a simple Gauss–Seidel iteration, which unfortunately does not converge here and is also too expensive. Krylov subspace methods are natural candidates but they do not work well without a good preconditioner. See, for instance, [32], [40] and [50] for multigrid accelerators. Another promising method is the full multigrid method. Yet another avenue of research is to reduce the ill-conditioning of the matrices. See [73] for work in this direction. Space-time ultraspherical methods ([53]) are worthy of investigation, as are space-time methods for delay differential equations.

Space-time methods are extremely robust methods which converge spectrally for most standard linear PDEs with standard boundary conditions. However without more sophisticated algorithms to speed up the linear algebra, space-time spectral methods are not faster than existing state-of-the-art algorithms.

Bibliography

[1] M. Al-Baali, *Descent property and global convergence of the fletcher-reeves method with inexact line search*, IMA Journal of Numerical Analysis **5** (1985), 121–124.

[2] N. Andrei, *Scaled conjugate gradient algorithms for unconstrained optimization*, Comput Optim Appl **38** (2007), 401–416.

[3] ———, *Accelerated adaptive Perry conjugate gradient algorithms based on the self-scaling memoryless BFGS update*, Journal of Computational and Applied Mathematics **325** (2017), 149–164.

[4] S. Babaie-Kafaki, *A note on the global convergence theorem of the scaled conjugate gradient algorithms proposed by Andrei*, Comput Optim Appl **52** (2012), 409–414.

[5] ———, *A modified scaling parameter for the memoryless BFGS updating formula*, Numerical Algorithms **72** (2016), 425–433.

[6] R. H. Bartels and G. W. Stewart, *Solution of the matrix equation $AX + XB = C$*, Commun. ACM **15** (1972), no. 9, 820–826.

[7] C. Bernardi and Y. Maday, *Spectral Methods, Techniques of Scientific (Computing (Part 2), Handbook of Numerical Analysis*, Elsevier, 1997.

[8] L. Blum, F. Cucker, M. Shub, and S. Smale, *Complexity and Real Computation*, Springer-Verlag, New York, 1998.

[9] J. P. Boyd, *Chebyshev and Fourier Spectral Methods*, 2nd ed., Dover, Mineola, New York, 2001.

[10] C. G. Broyden, *A class of methods for solving nonlinear simultaneous equations*, Math. Comp. **19** (1965), 577–593.

[11] C. G. Broyden, J. E. Dennis, and J. J. Moré, *On the local and superlinear convergence of quasi-newton methods*, IMA Journal of Applied Mathematics (Institute of Mathematics and Its Applications) **12** (1973), no. 3, 223–245.

[12] L. Brugnano, F. Iavernaro, and D. Trigiante, *Analysis of Hamiltonian boundary value methods (HBVMs): A class of energy-preserving Runge–Kutta methods for the numerical solution of polynomial Hamiltonian systems*, Communications in Nonlinear Science and Numerical Simulation **20** (2015), no. 3, 650–667.

[13] C. Canuto, M. Y. Hussaini, A. Quarteroni A., and Zang T. A., *Spectral Methods–Fundamentals in Single Domains*, Springer-Verlag, New York, 2006.

[14] A. J. Christlieb, R. D. Haynes, and B. W. Ong, *A parallel space-time algorithm*, SIAM Journal on Scientific Computing **34** (2012), no. 5, 233–248.

[15] P. G. Ciarlet, *Linear and Nonlinear Functional Analysis with Applications*, SIAM, Philadelphia, 2013.

[16] P. G. Ciarlet and C. Mardare, *On the Newton-Kantorovich theorem*, Analysis and Applications **10** (2012), no. 3, 249–269.

[17] X. Dai and Y. Maday, *Stable parareal in time method for first- and second-order hyperbolic systems*, SIAM Journal on Scientific Computing **35** (2013), no. 1, 52–78.

[18] Y. H. Dai and L. Z. Liao, *New conjugacy conditions and related nonlinear conjugate gradient methods*, Appl Math Optim **43** (2001), 87–101.

[19] Y. H. Dai and Y. J. Yuan, *A Nonlinear Conjugate Gradient Method with a Strong Global Convergence Property*, SIAM Journal on Optimization **10** (1999), no. 1, 177–182.

[20] D. W. Decker and C. T. Kelley, *Broyden's method for a class of problems having singular Jacobian at the root*, SIAM J. Numer. Anal. **22** (1985), no. 3, 566–574.

[21] J. E. Dennis, *On the convergence of Broyden's method for nonlinear systems of equations*, Math. Comp. **25** (1971), 559–567.

[22] J. E. Dennis and J. Moré, *A characterization of superlinear convergence and its application to quasi-Newton methods*, Mathematics of Computation **28** (1974), no. 126, 549–549.

[23] J. E. Dennis and J. J. Moré, *Quasi-Newton methods, motivation and theory*, SIAM Rev. **19** (1977), no. 1, 46–89.

[24] J. E. Dennis and R. Schnabel, *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*, SIAM, Philadelphia, 1996.

[25] A. Dutt, L. Greengard, and V. Rokhlin, *Spectral deferred correction methods for ordinary differential equations*, BIT Numerical Mathematics **40** (2000), no. 2, 241–266.

[26] R. D. Falgout, S. Friedhoff, T. V. Kolev, S. P. Maclachlan, and J. B. Schroder, *Parallel time integration with multigrid*, SIAM Journal on Scientific Computing **36** (2014), no. 6, 635–661.

[27] R. Fletcher, *Practical Methods of Optimization*, A Wiley-Interscience Publication, John Wiley & Sons, Ltd., Chichester, 1987.

[28] R. Fletcher and C. M. Reeves, *Function minimization by conjugate gradients*, The Computer Journal **7** (1964), no. 2, 149–154.

[29] B. Fornberg, *A Practical Guide to Pseudospectral Methods*, Vol. 360, Cambridge University Press, Cambridge, 1996.

[30] D. Funaro, *Spectral Elements for Transport-dominated Equations*, Springer-Verlag, Berlin, 1997.

[31] M. J. Gander, *50 years of time parallel time integration, multiple shooting and time domain decomposition methods*, Contrib. Math. Comput. Sci. **9** (2015), 69–113.

[32] M. J. Gander and S. Vandewalle, *Analysis of the parareal time–parallel time–integration method*, SIAM Journal on Scientific Computing **29** (2007), no. 2, 556–578.

[33] J. C. Gilbert and J. Nocedal, *Global Convergence Properties of Conjugate Gradient Methods for Optimization*, SIAM Journal on Optimization **2** (1992), no. 1, 21–42.

[34] G. H. Golub, S. Nash, and C. Van Loan, *A Hessenberg-Schur method for the problem* $AX + XB = C$, IEEE Trans. Automat. Control **24** (1979), no. 6, 909–913.

[35] D. Gottlieb and S. A. Orszag, *Numerical Analysis of Spectral Methods: Theory and Applications*, Society for Industrial and Applied Mathematics, Philadelphia, Pa., 1977.

[36] B. Y. Guo, *Spectral Methods and Their Applications*, World Scientific Publishing Co., Singapore, 1998.

[37] B. Y. Guo and Z. Q. Wang, *Legendre–Gauss collocation methods for ordinary differential equations*, Adv. Comput. Math. **30** (2009), no. 3, 249–280.

[38] J. S. Hesthaven, S. Gottlieb, and D. Gottlieb, *Spectral Methods for Time-dependent Problems*, Cambridge Monographs on Applied and Computational Mathematics, vol. 21, Cambridge University Press, Cambridge, 2007.

[39] R. A. Horn and C. R. Johnson, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, 1991.

[40] G. Horton and S. Vandewalle, *A space-time multigrid method for parabolic partial differential equations*, SIAM J. Sci. Comput. **16** (1995), no. 4, 848–864.

[41] L. V. Kantorovich, *Functional Analysis and Applied Mathematics*, Uspekhi Matematicheskikh Nauk **3(6)** (1948), no. 6(28), 89–185.

[42] T. Kato, *A Short Introduction to Perturbation Theory for Linear Operators*, Springer-Verlag, New York, 1982.

[43] C. Kou and Y. Dai, *A modified self-scaling memoryless Broyden–Fletcher–Goldfarb–Shanno method for unconstrained optimization*, Journal of Optimization Theory and Applications **165** (2015), no. 1, 209–224.

[44] A. S. Lewis and M. L. Overton, *Nonsmooth optimization via quasi-Newton methods*, Math. Prog. **141** (2013), 135–163.

[45] W. Liu, J. Sun, and B. Wu, *Galerkin-Chebyshev spectral method and block boundary value methods for two-dimensional semilinear parabolic equations*, Numerical Algorithms **71** (2016), 437–455.

[46] W. Liu, B. Wu, and J. Sun, *Space-time spectral collocation method for the one-dimensional Sine-Gordon equation*, Numerical Methods for Partial Differential Equations **31** (2015), no. 3, 670–690.

[47] Y. Liu and C. Storey, *Efficient generalized conjugate gradient algorithms, part 1: Theory*, Journal of Optimization Theory and Applications **69** (1991), no. 1, 129–137.

[48] S. H. Lui, *Numerical Analysis of Partial Differential Equations*, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2011.

[49] ———, *Legendre spectral collocation in space and time for PDEs*, Numerische Mathematik **136** (2017), 75–99.

[50] E. McDonald and A. Wathen, *A simple proposal for parallel computation over time of an evolutionary process with implicit time stepping*, Lect. Notes Comput. Sci. Eng., Springer, Cham (2016), 285–293.

[51] Y. Narushima and H. Yabe, *A survey of sufficient descent conjugate gradient methods for unconstrained optimization*, SUT Journal of Mathematics **50** (2014), no. 2, 167–203.

[52] J. Nocedal and S. J. Wright, *Nonlinear Optimization*, 2nd ed., Springer, New York, 2006.

[53] S. Olver and A. Townsend, *A fast and well-conditioned spectral method*, SIAM Review **55** (2013), no. 3, 462–489.

[54] S. S. Oren and D. G. Luenberger, *Self-Scaling variable metric (SSVM) alogorithmsS, Part I: Criteria and sufficient conditions for scaling a class of algorithms*, Management Science **4** (1974), no. 20, 845–862.

[55] S. S. Oren and E. Spedicato, *Optimal conditioning of self-scaling variable metric algorithms*, Mathematical Programming **10** (1976), 70–90.

[56] A. Perry, *A modified conjugate gradient algorithm*, Operations Research **26** (1978), no. 49, 1073–1078.

[57] M. J. D. Powell, *Some convergence properties of the conjugate gradient method*, Mathematical Programming **11** (1976), no. 1, 42–49.

[58] ———, *Nonconvex minimization calculations and the conjugate gradient method*, Lecture Notes in Math. **1066** (1984), 122–141.

[59] ———, *Convergence Properties of Algorithms for Nonlinear Optimization*, SIAM Review **28** (1986), no. 4, 487–500.

[60] L. Qi and J. Sun, *A nonsmooth version of Newton's method*, Math. Program **58** (1993), 353–367.

[61] S. C. Reddy and J. A. C. Weideman, *The accuracy of the Chebyshev differencing method for analytic functions*, SIAM J. Numer. Anal. **42** (2005), no. 5, 2176–2187.

[62] J. Shen, T. Tang, and L. L. Wang, *Spectral Methods*, Springer-Verlag, Berlin, Heidelberg, 2011.

[63] J. Shen and L. L. Wang, *Fourierization of the Legendre—Galerkin method and a new space–time spectral method*, Applied Numerical Mathematics **57** (2007), no. 5, 710 –720.

[64] Z. J. Shi and J. Shen, *Convergence of Liu–Storey conjugate gradient method*, European Journal of Operational Research **182** (2007), no. 2, 552–560.

[65] A. Solomonoff and E. Turkel, *Global properties of pseudospectral methods*, Journal of Computational Physics **81** (1989), no. 2, 239 –276.

[66] H. W. Sorenson, *Comparison of some conjugate direction procedures for function minimization*, Journal of the Franklin Institute **288** (1969), no. 6, 421–441.

[67] H. Tal-Ezer, *Spectral methods in time for hyperbolic equations*, SIAM Journal on Numerical Analysis **23** (1986), no. 1, 11–26.

[68] ———, *Spectral methods in time for parabolic problems*, SIAM Journal on Numerical Analysis **26** (1989), no. 1, 1–11.

[69] J. G. Tang and H. P. Ma, *Single and multi-interval Legendre τ-methods in time for parabolic equations*, Advances in Computational Mathematics **17** (2002), no. 4, 349–367.

[70] ———, *A Legendre spectral method in time for first-order hyperbolic equations*, Applied Numerical Mathematics **57** (2007), no. 1, 1 –11.

[71] T. Tang and X. Xu, *Accuracy enhancement using spectral postprocessing for differential equations and integral equations*, Commun. Comput. Phys. **5** (2009), no. 2, 779–792.

[72] L. N. Trefethen, *Spectral Methods in Matlab*, SIAM, Philadelphia, 2000.

[73] L. L. Wang, M. D. Samson, and X. Zhao, *A well-conditioned collocation method using a pseudospectral integration matrix*, SIAM J. Sci. Comput. **36** (2014), no. 3, 907–929.

[74] S. Wu and X. Liu, *Convergence of spectral method in time for Burgers' equation*, Acta Math. Appl. Sinica (English Ser.) **13** (1997), no. 3, 314–320.

[75] H. Yabe and M. Takano, *Global convergence properties of nonlinear conjugate gradient methods with modified secant condition*, Computational Optimization and Applications **28** (2004), no. 2, 203–225.

[76] J. C. Yakoubsohn, *A class of methods for solving nonlinear simultaneous equations*, J. Complexity **15** (1999), 239–281.

[77] N. Yamashita, *Sparse quasi-Newton updates with positive definite matrix completion*, Math. Prog. **115** (2008), 1–30.

[78] S. Yao, D. He, and L. Shi, *An improved Perry conjugate gradient method with adaptive parameter choice*, Numerical Algorithms **78** (2018), 1255–1269.

[79] L. Yi and Z. Wang, *Legendre Gauss type spectral collocation algorithms for nonlinear ordinary partial differential equations*, International Journal of Computer Mathematics **91** (2014), no. 7, 1434–1460.

[80] ———, *Legendre spectral collocation method for second-order nonlinear ordinary partial differential equations*, Discrete & Continuous Dynamical Systems **19** (2014), 299–322.

[81] W. Zhou and L. Zhang, *A nonlinear conjugate gradient method based on the mbfgs secant condition*, Optimization Methods and Software **21** (2006), no. 5, 707–714.

[82] G. Zoutendijk, *Nonlinear programming, computational methods*, Integer and Nonlinear Programming (1970), 37–86.