A posteriori error estimation for electronic structure calculations using ab initio methods and its application to reduce calculation costs

Yipeng Wang

SORBONNE UNIVERSITÉ
CRÉATEURS DE FUTURS DEPUIS 1257

SORBONNE UNIVERSITÉ
LJLL

Doctoral School **École Doctorale Sciences Mathématiques de Paris Centre**
University Department **Laboratoire Jacques-Louis Lions**

Thesis defended by **WANG YIPENG**
Defended on **December 18, 2023**
In order to become Doctor from Sorbonne Université
Academic Field **Mathématiques appliquées**
Speciality **Analyse numérique**

Estimation d'erreur a posteriori pour des calculs de structure électronique par des méthodes ab initio et son application pour diminuer le cout de calcul

A posteriori error estimation for electronic structure calculations using ab initio methods and its application to reduce calculation costs

Thesis supervised by Yvon MADAY

Committee members

Referees	Xiaoying DAI	Professor at Chinese Academy of Sciences	
	Simen KVAAL	Researcher at University of Oslo	
Examiners	Virginie EHRLACHER	Professor at Ecole des Ponts ParisTech	Committee President
	Antoine LEVITT	Junior Professor at Université Paris-Saclay	
	Andreas SAVIN	Research Director Emeritus at CNRS	
	Katharina SCHRATZ	Professor at SU	
	Martin VOHRALÍK	Research Director at INRIA	
Guest	Geneviève DUSSON	Researcher at CNRS	
Supervisor	Yvon MADAY	Professor at SU	

Colophon

Doctoral dissertation entitled “Estimation d’erreur a posteriori pour des calculs de structure électronique par des méthodes ab initio et son application pour diminuer le cout de calcul”, written by WANG Yipeng, completed on March 18, 2024, typeset with the document preparation system LaTeX and the yathesis class dedicated to theses prepared in France.

Keywords: electronic structure theory, coupled cluster method, numerical analysis, non-linear functions, error estimate, Gross-Pitaevskii equation, residual decomposition.

Mots clés : théorie de la structure électronique, méthode des clusters couplées, analyse numérique, fonction non linéaire, estimation d'erreur, équation de Gross-Pitaevskii, décomposition du residu.

This thesis has been prepared at

Laboratoire Jacques-Louis Lions
Sorbonne Université
Campus Pierre et Marie Curie
4 place Jussieu
75005 Paris
France

☎ +33 1 44 27 42 98
Web Site `https://ljll.math.upmc.fr/`

To my dearest family: my grandparents, my parents, my sister and my two-year-old cute niece.

回首向来萧瑟处，归去，也无风雨也无晴。

宋 苏轼

Estimation d'erreur a posteriori pour des calculs de structure électronique par des méthodes ab initio et son application pour diminuer le cout de calcul
A posteriori error estimation for electronic structure calculations using ab initio methods and its application to reduce calculation costs

Abstract

The thesis is concerned with the error analysis of electronic structure calculation. The long term goal is to, in one hand, derive computable a posteriori error estimator for ab initio methods and, in the other hand, propose near-optimal computational cost strategy for the numerical calculation of those methods based on the a posteriori error estimation and the separation of the discretization and iteration error sources. In the first part of the thesis, we introduce a new well-posedness analysis for the single reference coupled cluster method based on the invertibility of the CC derivative. Under the minimal assumption that the sought-after eigenfunction is intermediately normalisable and the associated eigenvalue is isolated and non-degenerate, we prove that the continuous (infinite-dimensional) CC equations are always locally well-posed. Under the same minimal assumptions and provided that the discretization is fine enough, we prove that the discrete Full-CC equations are locally well-posed, and we derive residual-based error estimates with guaranteed positive constants. The second part of the thesis focus on the application of a posteriori error estimation to construct near-optimal path when approximating the solution of PDEs. We firstly apply a probabilistic method to explore an optimal path that minimizes the cost for the numerical resolution of linear and nonlinear elliptic source problems. Based on the analysis of those optimal paths, we propose two near-optimal strategies to achieve a given accuracy based on the error sources decomposition of the error estimator. Finally, we validate the feasibility of those near-optimal strategies by applying them to the numerical approximation of a nonlinear eigenvalue problem, i.e., the Gross-Pitaevskii equation.

Keywords: electronic structure theory, coupled cluster method, numerical analysis, non-linear functions, error estimate, Gross-Pitaevskii equation, residual decomposition.

Résumé

La thèse porte sur l'analyse des erreurs dans le calcul de la structure électronique. L'objectif à long terme est, d'une part, de dériver un estimateur d'erreur a posteriori calculable pour les méthodes ab initio et, d'autre part, de proposer une stratégie de coût de calcul quasi-optimale pour le calcul numérique de ces méthodes basée sur l'estimation d'erreur a posteriori et la séparation des sources d'erreur de discrétisation et d'itération. Dans la première partie de la thèse, nous introduisons une nouvelle analyse de bien posé pour la méthode de cluster couplé à référence unique basée sur l'inversibilité de la dérivée CC. Sous l'hypothèse minimale que la fonction propre recherchée est normalisable de façon intermédiaire et que la valeur propre associée est isolée et non dégénérée, nous prouvons que les équations CC continues (en dimension infinie) sont toujours bien posées localement. Sous les mêmes hypothèses minimales et à condition que la discrétisation soit suffisamment fine, nous prouvons que les équations CC discrètes sont localement bien posées, et nous dérivons des estimations d'erreur basées sur les résidus avec des constantes positives garanties. La deuxième partie de la thèse se concentre sur l'application de l'estimation d'erreur a posteriori pour construire un chemin quasi-optimal lors de l'approximation de la solution d'EDP. Nous appliquons d'abord une méthode probabiliste pour explorer un chemin optimal pour la résolution numérique de problèmes elliptiques linéaires et non linéaires en minimisant le coût de calcul. Sur la base de l'analyse de ces chemins optimaux, nous proposons deux stratégies quasi-optimales pour atteindre une précision donnée, basées sur la décomposition des sources d'erreur de l'estimateur d'erreur. Enfin, nous validons la faisabilité de ces stratégies quasi-optimales en les appliquant à l'approximation numérique du problème des valeurs propres, c'est-à-dire l'équation de Gross-Pitaevskii.

Mots clés : théorie de la structure électronique, méthode des clusters couplées, analyse numérique, fonction non linéaire, estimation d'erreur, équation de Gross-Pitaevskii, décomposition du residu.

Laboratoire Jacques-Louis Lions
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France

Contents

0.1 Acknowledgments

I would like to express my sincere gratitude to the individuals who played a pivotal role in the completion of this doctoral dissertation.

This thesis has been particularly exciting thanks to the collaboration of Yvon Maday and Muhammad Hassan, to whom this work owes a great deal.

I extend my heartfelt appreciation to the peer reviewers, Doctor Simen Kvaal and Professor Xiaoying Dai, whose insightful comments and constructive feedback significantly enhanced the quality of this work. Their expertise and careful review were instrumental in shaping the final version of this dissertation. I am deeply indebted to the members of my dissertation committee, Geneviève Dusson, Virginie Ehrlacher, Antoine Levitt, Andreas Savin, Katharina Schratz and Martin Vohralík. Their scholarly guidance, valuable suggestions, and dedicated time invested in the evaluation of this dissertation have been invaluable.

I would like to say thank you to Monsieur Bertrand MERCIER. Without his recommendation, I would not have the chance to meet and work with my current doctoral supervisor. Then I'd like to express my thanks to my supervisor, Yvon MADAY. Many thanks to all your help and guidance. Your precise intuition and ability to translate some ideas into mathematical proofs is so impressive! I really learned a lot during the time I worked with you. In fact, I would never have imagined that I will do a mathematical thesis in Paris. As an engineer student, my normal life path would have been to get a PhD in industry. So it's a huge surprise (and pleasure) that I came to spend several years in such a top applied math lab with so many excellent professors and PhD students. I really learned a lot here, from Yvon, from my colleagues and from my friends. I'm not so sure about my future career, but I will never regret my decision to spend several years on my Ph.D. in LJLL where I met so many kind people.

Sincere thanks to Hassan for all the favors. As an engineer student, in fact I didn't know how to become a math doctor. I just followed the same idea of any engineer student: I receive a project. Then for resolving a problem, we do some research. We do research to determine what we need to know but we don't go into the details of those results. We just apply them to solve the practical problem. However, the math research is quite different from that. This is what I learned from Hassan and Yvon: we give all details of our research. We define everything properly and derive our results rigorously. We pay attention to what we say, how to express rigorously an idea, how to present what we found at the blackboard during a discussion, etc. Hassan, you are a great example of what a young doctor in math should look like and I really learned a lot from you. In addition, thanks a lot for teaching me how to write a math article. Without the help from you and Yvon, it would never been possible for me to finish the writing of my thesis.

Moreover, I would like to express my sincere thanks to Antoine Levitt, Eric Cancès and Mi-song Dupuy for their invaluable insights and discussions. Additionally, thanks to Mi-Song, Emmanuel, Thomas, Julien, Cindy and Fabrice for organizing the EMC2 seminar, the GT for electronic structure computation and quantum computing. I have learned a lot from all these talks and met a lot of excellent researchers. I believe that all these experiences will help me a lot in my future career!

Thanks to Madame Catherine Drouet, who was in charge of the administrative process for my PhD admission. It's a pity that you're already retired so I can't thank you in person for all the help you have given me. In addition, thanks to Malika, Erika, Salima and Corentin for all the help and assistance on the administrative side. Thanks to Khashayar, Pierre-Henri and Antoine for all the IT support including the use of the server HPC2, which is of essential importance to my PhD work. Here I also want to thank to Professors Bruno Després and Marie Postel for lending me their key when I forgot my key and got locked out of my office. Last but not least, I really want to mention that I would never had imagined that I will make friends with a professor

here. Miraculously, I did. In fact, I'm quite surprised that Professor François Murat works very late in the evening. So sometimes when I worked also very late, before leaving I went to the office of François to say 'Au revoir, bonne soirée!' such that he didn't feel so quiet in the corridor. (Honestly, I really feel a bit lonely if I work alone at 21 o'clock in my office, so I guess maybe it's the same for others and it's better to make some sound?) Later, it became a habit and I didn't think that I could have made a friend in this way. Whatever, thanks to François to let me know that I'm not the only one working at night here, which makes me feel much better.

Thanks to Elise and Anne-Françoise, I learned from you how to work with Yvon and I won't forget the moment that we were all lined up in front of the conference room. Thanks to Jules for all the help that you gave to me and other PhD students in our lab. You are really nice, kind and responsible.

Thanks to all the GTT organizers: Jules, Allen, Lise, Matthieu, Noemi, Pierre, Maria, Charles, Guillaume, Lucas, Ludovic and Zhe for organizing the GTT every week. Especially for Lise and Matthieu, it has been a pleasure working with you. Meanwhile, I'd like thank all the organizers of the Thé du labo during the past four years: Elise, Agustin, Anatole, Jesus, Ramon, Rui, Chourouk, Ruikang and Eleanor. Thank you for all your efforts to keep our lab a happy and lively place. In fact, there are many activities in our lab and thanks to all our colleagues who have made their efforts to make our lab better: Thanks to Giorgia, Rémi, Sylvain, Antoine, Emma, Juliette, Lucas, Roxane, Robin, Thomas to work as representatives of the PhD students and help us to solve many problems including the thesis work. Thanks to Elena and Yvonne for organizing the First Year Welcome Day. Thanks to Ioanna for organizing Infomath. Thanks to Liangying, Nicolai, Mingyue to help arranging the office. Thanks also for preparing posters of PhD students (Sorry, here I would like to mention the name, but I really don't know who make it except Fatima for the year 2019.)

Next, I want to express my thanks to my office mates. When I arrived in 2019, there were Noemi, Rémi, Christophe, Élise, Alex and Gabriela. It was a really good (and unfortunately too short) time to work with you in the same office. The atmosphere in the office was very relaxed and active. Unfortunately, later there was Covid and then I went back to China. When I returned to France in the end 2021, it has changed a lot: Christophe and Alex have left, Gabriela and Elise defended within a short period of time, Noemi went to England for a visit before defending and only Rémi and me were left in the office. Several months later, Rémi and Noemi left, too. I was the only one staying in our office and keeping memories of our time together. Now I will leave too and this part of the history about office 15-25 326 finally ends. Now in this office, we have new members: Anna, Cristobal, Houssam, Valentin, Nathalie, Mingyue, Gala and Sebastian. Thank you for your company and it's my pleasure to have so many good office mates, sincerely.

I would like also to say thanks thanks thanks to our Crous team, which is so important so me!!! (Ok, I admit that it's Crous that is important to me, but I'm also grateful that I can eat with you every workday!) Thank you for the organizers of Crous team, especially memes sent by Pierre. Because I need to take a nap in the noon, I couldn't joint the happy cafe time. It's a pleasure to share our ideas and histories in Crous or during the Thé du lab. I wish you 'Bon Appetit' every day.

Thank you to Chufa, one of my best friends, thanks you for all our mobile gaming time. I'm happy that there is someone who plays video games with me in the weekend (before that you returned back to China). Thanks to Gong for sharing the information about 'how to find Yvon' and 'Is Yvon here today'. And thanks, my IFCEN doctoral classmates in France, for all moments that we shared together. Thanks to Anatole for your excellent magic show. Thanks to Roxane for inviting me to the lunch of your anniversary. Thanks to my chinese colleagues in our lab: Siyuan, Xinran, Shengquan, Chaoyu, Yangyang, Shijie, Allen, Rui, Jingrui, Haibo, Liangying, Boxi, Ruiyang, Zeyu, Siguang, Jingyi, Yue, Ruikang, Mingyue, Zhe, zhengping. Thank you for

all the Friday dinners and al the happy moments that we shared together.

By the way, thanks to Chat GPT 3.5 which plays the role of a very good English writing teacher.

My world is very small: my home in China, my little studio in Villejuif, LJLL where I work, the subway line 7 that I take every workday, the supermarkets where I go to every weekend (maybe with the addition of Trantranzai?). That's all of my life.

For my french colleagues. In fact, every time you went back to our lab, I'm happy to see you again. I don't know what to say but I'm really happy to see you again. I'm not the type of person who can easily find a conversation topic that flows freely, especially if I need to speak in French or English. In most times, I just stand by and listen to you guys talk and try to understand the conversation, feeling like time turns back and we are all still undergraduate PhD students. However, when there is a meeting, there is a parting. Now it's my turn to leave. Maybe one day I will come back. I imagine that I will sit by quietly as I do now, watching the persons I know and I don't know talk and trying to understand their conversation.

Thank you to the people who have appeared in my life, my family, teachers, classmates, friends, thank you for being there and making my life so fulfilling!

And finally here is the final conclusion: This dissertation would not have been possible without the collective contributions of these individuals and institutions. I am truly grateful for the privilege of working with such remarkable people.

Chapter 1

Introduction

Computational quantum chemistry is by now widely regarded as one of the central pillars of modern chemistry. In quantum chemistry, the behavior of matter is described using wave functions and governed by a so-called Hamiltonian operator acting on a Hilbert space of these wave functions. In the so-called non-relativistic Born-Oppenheimer setting, the nuclei of the molecule under study are treated as clamped, point-like particles. The goal is to study the evolution of the electrons, a field of study called electronic structure theory. In electronic structure calculation, the primary difficulty is the extremely high-dimensionality of solution space, which depends exponentially on the size of the electronic system. For a system containing Q electrons, the sought-after ground state (i.e., the lowest eigenfunction) of the electronic Hamiltonian depends on $3Q$ spatial variables. A naive application of traditional numerical methods such as finite element approximations or spectral schemes, etc., therefore fails spectacularly. After nearly a century's development, a number of ab initio (first principles-based) deterministic numerical methods for approximating the ground state energy have been constructed. However, even these methods are well established, many questions pertaining to the mathematical error analysis of these methods remain unconsidered.

Error estimation is usually classified into two types: the *a priori* error estimation and the *a posteriori* error estimation. *A priori* error estimates typically have the following form:

$$\|u - u_N\| \leqslant CN^{-k}, \tag{1.1}$$

where u is the exact solution to our problem and u_N is the approximate solution in a discretisation space depending on a parameter $N > 0$ measuring the number of degrees of freedom of this discretization space, $\|\cdot\|$ is a certain norm, and C, k are positive constants associated with the measure of the regularity of u. Estimate (1.1) ensures that the approximation error $\|u - u_N\|$ goes to zero as N goes to infinity. In addition, the parameter k indicates the convergence rate of the discrete solutions towards the exact one. The disadvantage of the above estimate is that the constant C is, in general, unknown because it is a function of the unknown solution $C = C(u)$. Conversely, *a posteriori* error estimates typically have the following form:

$$\|u - u_N\| \leqslant P(u_N, C), \tag{1.2}$$

where P is a function of the known numerical solution u_N and a set of known or computable data C at a cost lower or similar to the cost spent to determine u_N. Estimate (1.2) provides computable error bounds controlling the difference between our approximate result and the true solution. In numerical approximation methods such as finite element method (FEM) [109, 6, 34,

108, 80] or finite difference method(FDM) [69, 64, 76, 73], it may also provide further information on the error in different region of the computational domain. This is the case when the bound $P(u_N, C)$ appears as a sum of contributions related to a small region of the computational domain. We can then refine the computational domain where the local error is relatively high. The estimator is then named 'indicator' and the method that stems is known as the adaptive mesh method.

On the one hand, the long term goal of our work is to derive computable a posteriori error estimator for ab initio methods and, on the other hand, propose near-optimal computational cost strategies for numerical calculations of these electronic structure methods. For the error analysis portion of this work, we introduce a new well-posedness analysis for the single reference coupled cluster method and we derive residual-based error estimates with guaranteed positive constants. With regards to improving calculation accuracy with limited resources, we propose two near-optimal strategies to achieve a given accuracy for the numerical solution of Gross-Pitaevskii type equations.

In the following sections, we will introduce more in detail into these aforementioned topics and provide an overview of the content in each chapter.

1.1 The coupled cluster equations

The starting point for a derivation of the coupled cluster (CC) equations is the time-independent Schrödinger equation: For a system containing Q electrons and M nuclei, under the Born-Oppenheimer approximation, the Hamiltonian operator that describes the electronic behavior is given by

$$H := -\frac{1}{2}\sum_{j=1}^{Q}\Delta_{\mathbf{x}_j} + \sum_{j=1}^{Q}\sum_{n=1}^{M}\frac{-Z_n}{|\mathbf{z}_n - \mathbf{x}_j|} + \sum_{j=1}^{Q}\sum_{i=1}^{j-1}\frac{1}{|\mathbf{x}_i - \mathbf{x}_j|}, \tag{1.3}$$

where $\{Z_n\}_{n=1}^{M}$ and$\{\mathbf{z}_n\}_{n=1}^{M}$ are the charges and positions of these M nuclei respectively and $\{\mathbf{x}_i\}_{i=1}^{Q}$ are the positions of the Q electrons. The behavior of the electrons is determined by the eigenvalues of the Hamiltonian operator H, i.e.,

$$H\Psi^* = \mathcal{E}^*\Psi^*. \tag{1.4}$$

Here the eigenvector Ψ^* is the wave function describing the state of the electronic system while the eigenvalue $\mathcal{E}^*$ corresponds to the energy of the system under state Ψ^*. Of particular importance in electronic structure calculations is the lowest eigenvalue of H, which is the ground state energy of the system $\mathcal{E}^*_{\mathrm{GS}}$ with corresponding state Ψ^*_{GS}.

It is well-known that electrons are Fermionic particle and it obeys the so-called Pauli-exclusion principle. In mathematical term, this means that the wave function Ψ is antisymmetric, i.e.,

$$\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots \mathbf{x}_Q) = -\Psi(\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots \mathbf{x}_Q) \qquad \forall\, i, j \in \{1, \ldots, Q\} \text{ with } i \neq j.$$

In order to compute a solution to Equation (1.4), it is necessary to approximate the antisymmetric wave function. The simplest idea is the separation of variables approach, i.e., to write the wave function as a product of Q functions, each of which only depends on one space variable $\{\mathbf{x}_i\}_{i=1}^{Q}$. Unfortunately, following this idea, the resulting wave function is not antisymmetric. In 1926, Heisenberg [50] and Dirac [36] independently proposed to use Slater determinants to represent the wave function, these Slater determinants being functions that trivially satisfy the antisymmetric property. Given Q ortho normal functions of one space variable $\{\phi_i(\mathbf{x})\}_{i=1}^{Q}$ (which we also call

single particle orbitals), the Slater determinant Φ constructed from $\{\phi_i(\mathbf{x})\}_{i=1}^{Q}$ is defined as

$$\Phi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \det\big(\phi_i(\mathbf{x}_j)\big)_{i,j=1}^{Q},$$

with

$$\int_{\mathbb{R}^{3Q}} |\Phi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q)|^2 d\mathbf{x}_1 d\mathbf{x}_2 \ldots d\mathbf{x}_Q = 1. \tag{1.5}$$

A standard Galerkin discretisation strategy involves approximating the antisymmetric wave functions by selecting a finite number of single particle orbitals $\{\phi_i\}_{i=1}^{K}$ $(K > Q)$ and constructing a finite number of Slater determinants using any Q non identical orbitals. Then the sought-after ground state wave function Ψ_{GS}^{*} is expressed as a linear combination of those Slater determinants and Equation (1.4) reduces to a linear system of equations determining unknown coefficients appearing in the linear expansion of these Slater determinants. This is known as the full configuration interaction (Full CI) method.

For practical calculations, the full configuration interaction approximate space is still too large, i.e., with dimension being $\binom{K}{Q}$. Therefore, furthermore simplification must be made to balance the solution accuracy and computational costs. The fundamental idea of such approaches is as follows: By selecting a reference determinant Ψ_0, any other Slater determinant can be viewed as replacing n $(1 \leqslant n \leqslant Q)$ single particle orbitals by n new orbitals. Conventionally, we call these Q orbitals used to construct Ψ_0 *occupied orbitals* and the remaining orbitals unoccupied or virtual. In the mathematical term, we describe this orbitals replacement action as an excitation operator acting on Ψ_0. In addition, we can classify these so-called excited Slater determinants according to the number of orbital replacements, i.e., single excited Slater determinant, double excited Slater determinant, triple excited Slater determinant, etc. The space is then restricted to contain Ψ_0 and only specific orders of excited Slater determinants, e.g., the CISD method containing Ψ_0 and only single and double excited Slater determinants.

One common problem of such truncated CI methods is that they do not satisfy the size consistency requirement, which is a major concern in electronic structure calculations. The size-consistency property concerns the additive separability of the energy of a molecular system, i.e., if we divide a molecular system into two non-interacting subsystems (e.g. separated by an infinite distance), then the energy of the whole system computed by a size-consistent method is the sum of the energies of these two subsystems taken separately. Compared to the linear parameterization used in the truncated CI methods, the coupled cluster (CC) method uses an exponential parameterization, which fulfills the size consistency condition for the truncated versions of the coupled cluster method under the assumption that the reference determinant is multiplicatively separable (see e.g., the work of Andreas Savin in [86] for the meaning of this term).

For any Φ in the Galerkin approximate space that satisfies the so-called intermediate normalisation condition

$$\int_{\mathbb{R}^{3Q}} \Phi(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q)\Psi_0(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) d\mathbf{x}_1 d\mathbf{x}_2 \ldots d\mathbf{x}_Q = 1, \tag{1.6}$$

the component $\Phi - \Psi_0$ can be expressed as a linear combination of excited Slater determinants. As each excited Slater determinant can be expressed as an excitation operator acting on Ψ_0, the component $\Phi - \Psi_0$ can also be viewed as a weighted summation of excitation operators, also known as a cluster operator acting on Ψ_0, i.e., $\mathcal{S} = \sum_{\mu \in \mathcal{J}} \boldsymbol{s}_\mu \mathcal{X}_\mu$ where $\{\mathcal{X}_\mu\}_{\mu \in \mathcal{J}}$ and $\{\boldsymbol{s}_\mu\}_{\mu \in \mathcal{J}}$ are a sequence of excitation operators and corresponding coefficients respectively and $\mathcal{J}$ is an index set of excitation operators. It is known [87] that for intermediate normalised wave function Φ

there exists a unique cluster operator $\mathcal{T} = \sum_{\mu\in\mathcal{G}} \boldsymbol{t}_\mu \mathcal{X}_\mu$ such that

$$\Phi = e^{\mathcal{T}}\Psi_0. \tag{1.7}$$

In this manner, a linear combination of basis vectors is rewritten as an exponential of cluster operator acting on Ψ_0. Inserting the above exponential parameterization for Ψ^* into Equation (1.4) yields the so-called single reference coupled cluster (CC) equations

$$\forall \mu \in \mathcal{G}, \quad \left\langle \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}^*} H e^{\mathcal{T}^*} \Psi_0 \right\rangle = 0. \tag{1.8}$$

A solution of problem (1.8) is a sequence of coefficients $\mathbf{t}^* = \{\mathbf{t}_\mu\}_{\mu\in\mathcal{G}}$ such that $\mathcal{T}^* = \sum_{\mu\in\mathcal{G}} \boldsymbol{t}_\mu \mathcal{X}_\mu$. After obtaining $\mathbf{t}^*$, the corresponding wave function is given by $\Psi^* = e^{\mathcal{T}^*}\Psi_0$ and the energy is calculated via

$$\mathcal{E}^* := \left\langle \Psi_0, e^{-\mathcal{T}^*} H e^{\mathcal{T}^*} \Psi_0 \right\rangle. \tag{1.9}$$

Similar to CI methods, only the truncated CC methods like the CCSD or CCD methods are widely used in practical calculation. And one of the major interest of truncated CC methods is that the size consistency condition still holds [87] under the assumption regarding the reference determinant presented earlier. In addition, the truncated coupled cluster approximations could be viewed as classical Galerkin discretisation of the so-called continuous coupled cluster equations resulting in a set of nonlinear equations.

Our work aims at proving the local well-posedness of both the continuous CC equations and its Galerkin discretisations. We view the coupled cluster equation as a root-finding problem for a nonlinear function and we attempt to show the local existence and uniqueness of the roots $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{G}}$ of the coupled cluster function.

The mathematical analysis of the coupled cluster equations was begun by Reinhold Schneider and Thorsten Rohwedder in a series of three papers [87, 82, 83]. It is shown that under some assumptions, the non-linear coupled cluster function is *locally, strongly monotone* which can be exploited to prove the local well-posedness of both the continuous CC equations and its Galerkin discretisations. They also derived optimal error estimates for the coupled cluster energies using the dual-weighted residual approach of Rolf Rannacher and co-workers [92]. Unfortunately, establishing the local monotonicity property requires rather pessimistic assumption on the closeness of the targeted root $\boldsymbol{t}^*$ to 0. Essentially, the local well-posedness analysis and the resulting error estimates only hold in a perturbative regime $\boldsymbol{t}^* \approx 0$. On the other hand, as we discuss in more detail in Remark 2.4.2 in Chapter 2.4, in many practical situations where the CC method is *known numerically* to yield accurate approximations, the sought-after root $\boldsymbol{t}^*$ is *not* in the perturbative regime. For such problems, the existing a priori analysis yields estimates with *negative* constants. The *a priori* analysis having failed, there is also no hope of developing *a posteriori* error estimates for practical coupled cluster simulations which will be the ultimate goal of our numerical analysis.

The contribution of our work is to develop a new *a priori* error analysis for the single reference coupled cluster equations that is valid under more general conditions. The analysis we present here– motivated by the existing literature on non-linear numerical analysis (see, for instance, [12, 96])– is based on the invertibility of the Fréchet derivative of the non-linear coupled cluster function, which is established using a classical inf-sup-type approach. Compared to the local monotonicity approach, such an inf-sup condition seem to hold under more general assumptions.

In Chapter 2.4, we show that under the assumption that the sought-after eigenvalue $\mathcal{E}^*$ of the electronic Hamiltonian H is non-degenerate and that the corresponding eigenvector Ψ^* is intermediately normalisable, then the continuous CC solution $\boldsymbol{t}^*$ is unique in its neighborhood

with the existence of $\boldsymbol{t}^*$ being deduced directly from the assumption.

In Chapter 2.5, we consider the discretisation of the continuous CC equations, namely, the Full-CC equations in a finite basis. We show that under similar assumptions, these discrete equations are also locally well-posed in the asymptotic limit.

In this work, we focus on the continuous (infinite-dimensional) CC equations and a specific version of the discrete CC equations, namely, the Full-CC equations in a finite basis. The extension of our analysis to more general discretisation (the so-called *truncated* CC equations [52, Chapter 13]) is a work in progress.

1.2 The optimal path problem and its application

The aim of developing *a posteriori* error estimators is not only to give computable error bounds, but also to improve the calculation accuracy. *A posteriori* error analysis was initially developed for finite element methods (FEM). In FEM, the computational domain is discretized by a mesh. The *a posteriori* error analysis may lead to estimate the error in different regions of the mesh. Based on the information provided by the *a posteriori* error estimator, we can refine the mesh in regions where the error is relatively high or coarsen the mesh where the error is relatively small. In such a way, computational costs are kept under control while accuracy are still maintained. This method is known as the adaptive mesh method [100, 7, 72].

For other approximation methods like the planewave (Fourier) approximations studied in this work, the mesh (collocation) is regular and directly in relation with the number of Fourier modes. This is the sense in refining it locally. The application of *a posteriori* error estimation as an adaptive strategy to improve the calculation accuracy is thus only global and the idea of improving calculation accuracy thus consists of including gradually more Fourier modes in the numerical solution process.

In our work, the numerical error is divided into two parts which originate respectively from the limited degrees of freedom of the discretization space and the limited number of iterations in the numerical solution process of the discrete non linear problem. For obtaining the approximate solution achieving target accuracy, a straightforward approach is to first fix a large enough discretization space and then perform iterative calculations, hoping that after enough number of iterations the final solution will be accurate enough. In this case, the discretization error is fixed at the beginning and only the iteration error is reduced during the calculation process. Of course, the final accuracy is guaranteed but the cost is a much longer computation time or much more practically consumed computational resources. In such a case, in order to achieve a target accuracy while minimizing the computational costs, the reduction of the discretization and iteration errors should be carefully synchronized, this is the idea adopted in our work.

The first step of our work is to explore the best error balance strategy such that the computational costs are minimized while guaranteeing a final calculation accuracy. This portion of the work is called the optimal path problem and is shown in Chapter 4.3. We name 'path' any possible calculation process that allows to reach the target tolerance:

1. We firstly fix a discretization space and perform several iterations in this space.
2. We continue the calculation in a larger discretisation space with several more iterations.
3. We repeat this process several times until we finally obtain a solution satisfying the accuracy requirement.

A path collects the information about the choice of discretization spaces and number of iterations performed in corresponding discretization spaces and outputs it as an array. We explore the

optimal path which minimizes the computational costs using the probabilistic *threshold accepting* (TA) method, which is a variant of the well-known *simulated annealing* (SA) method [58], with a simpler structure (see Chapter 4.3).

In this work, we focus on the numerical resolution of Gross-Pitaevskii type equations, a simple toy but representative problem in quantum chemistry. In the first place, we solve numerically a simple linear elliptic source problem. For a given accuracy, we apply the TA method to obtain the optimal path. Displayed in Figure 1.1 is result for a specific chosen of model parameters. There are two paths in Figure 1.1. The blue line represents the calculation process performing 13 iterations in space $N = 100$. This is the straightforward but expensive way to perform the calculation. The red line represents the optimal path, which consists of performing seven iterations for $N = 3$, four iterations for $N = 4$, one iteration for $N = 6$ and finally one iteration for $N = 100$. Compared with the fixed $N = 100$ path, the optimal path cuts down the total cost of computation by a factor thirteen. From Figure 1.1, we observe that after jumping from $N = 6$ to $N = 100$, only 1 iteration for $N = 100$ is sufficient to achieve the goal accuracy, which also means that the discretization error gap is easily covered by one iteration after enlargement of discretization space.

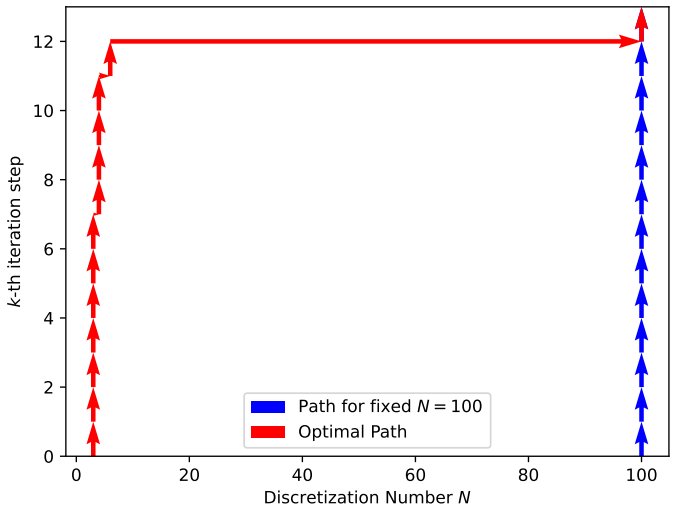


Figure 1.1: The optimal path for the linear case

The second step of our work is to explore the mechanism generating the two types of optimal paths (i.e., for the linear and nonlinear cases) and this part of work is presented in Chapter 4.4. We perform several complementary simulations and compare the results to unlock the key factor in the iterative scheme determining the form of optimal paths. Based on this discovery, we give a detailed analysis of the convergence rates of iterative schemes including general iteration processes and the specific iteration after the enlargement of discretization space. With additional numerical verification, we confirm the mechanism generating such two types of optimal paths.

Based on our understanding of such optimal path, we next develop simple and computationally cheap strategies that can produce near-optimal paths and this part of our work is presented in Chapter 4.5. To do so, we apply residual-based *a posteriori* error estimation for both of the two source problems and decompose the total residual into so-called discretisation and iteration

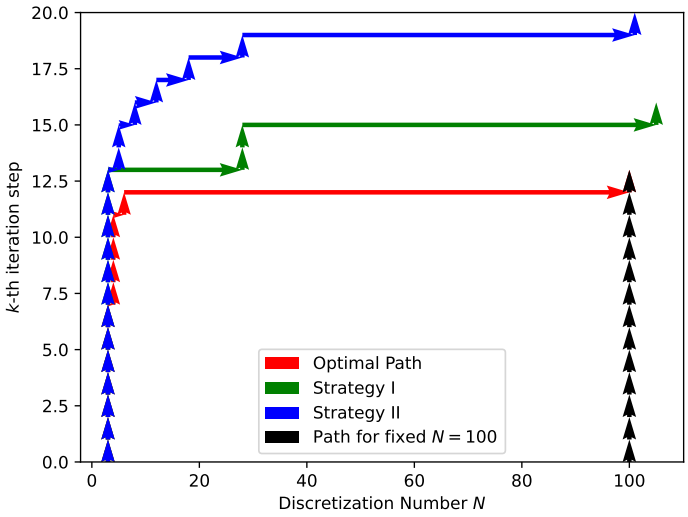


Figure 1.2: Comparison between optimal path and nearly optimal strategies

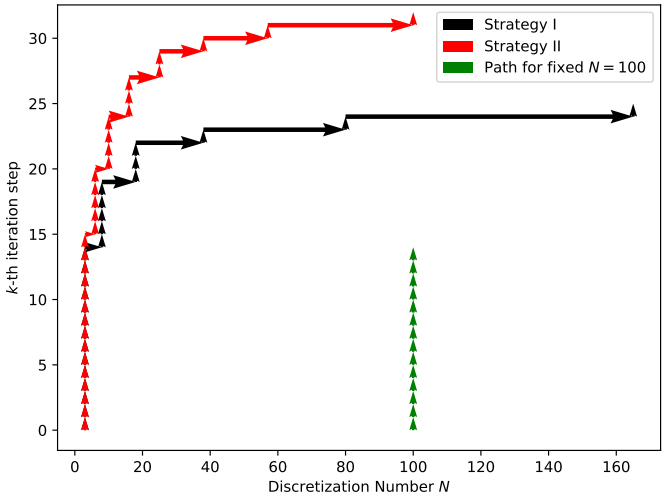


Figure 1.3: Nearly optimal strategies for the eigenvalue problem

residuals, each of which represents a bound of the corresponding source of error. An important feature of our residual-based estimator is that we can use it to predict both the iteration and discretisation errors after the enlargement of the discretisation space thus allowing us to select an appropriate choice of N for the next iteration. In this case, we can compare the costs and predict the benefits (the decrease of the residual) in order to pick near-optimal discretization numbers N in the path. Using these features, we propose two near-optimal strategies and compare the

resulting near-optimal paths with the optimal one. We show in Figure 1.2 the result of the linear case as an example. It is readily seen that the computational cost of the near-optimal paths is very close to the optimal one.

The last step of our work is to apply these two near-optimal strategies to the numerical solution of eigenvalue problem, the Gross-Pitaevskii equation. This part of work is presented in Chapter 5. We compare the differences between numerical solutions of the source and this new eigenvalue problem. We check that the mechanism controlling the form of the original optimal path in the source problem is also in play in the eigenvalue problem. Finally, we apply the same near-optimal strategies to the eigenvalue problem. A representative result is shown in Figure 1.3.

While this thesis involves work only on the nonlinear Gross-Pitaevskii-type equations, the long term goal is to propose near-optimal strategy for the numerical solution of more complex quantum chemistry calculation method such as the Hartree-Fock equations, the Khon-Sham eqautions, eventually other ab initio methods.

Part I

Analysis of the single reference coupled cluster method for electronic structure calculations

Chapter 2

The coupled cluster equations

This chapter presents the results of [49]. This work was carried out in collaboration with Yvon Maday and Muhammad Hassan.

2.1 Introduction

Computational quantum chemistry is by now widely regarded as one of the central pillars of modern chemistry, as evidenced by the award of two Nobel prizes (Walter Kohn and John Pople (1998) [94]; Martin Karplus, Michael Levitt, and Arieh Warschel (2013) [95]) in recent years. The field is typically thought to have begun with the pioneering work of Walter Heitler and Fritz London [51] in the 1920s but major, concurrent advances were due to Vladimir Fock, Douglas Hartree, Egil Hylleraas and John Slater [48, 44, 91, 90, 57, 56] among others. These first developments were followed by seminal contributions in the post-war period by the likes of Francis Boys [8], Jiří Čížek [25], George Hall [47], Clemens Roothan [85, 84] and many others (see, for instance, [15, Chapter 1.2] for a more comprehensive account). The subsequent explosion in available computing resources which began in the 1970s helped spur a tremendous development in the field (see, e.g., the development of computer software such as POLYATOM [5], IBMOL [26], and GAUSSIAN 70 which is still used today [46]), and quantum-chemical simulations are today routinely performed by thousands of researchers, complementing painstaking laboratory work on the design of new compounds for *sustainable energy*, *green catalysis*, and *pharmaceutical drugs* (see, e.g., [35, 65, 68, 33, 53] and the references therein). Indeed, according to the 2021 annual report of the European High Performance Computing Joint Undertaking (EuroHPC-JU), nearly a quarter of of the simulations running on the supercomputers Karolina and Vega pertained to chemical and material science simulations, with similar or higher numbers reported by supercomputing centers in Germany [41], Italy [24], and Switzerland [2].

The goal of quantum chemistry is to obtain a quantitative description of the behaviour of matter at the atomic scale, i.e., when matter is viewed as a collection of nuclei and electrons. In the so-called non-relativistic Born-Oppenheimer setting, the nuclei of the molecule under study are treated as clamped, point-like particles, and the aim is to describe the evolution of the electrons in the effective electrostatic potential generated by the static configuration of positively charged nuclei, this field of study being known as electronic structure theory. The behaviour of the electrons in this situation is governed by the spectrum of the so-called electronic Hamiltonian– a semi-unbounded, self-adjoint operator acting on an L^2-type Hilbert space of antisymmetric functions. It has been known since the seminal work of Grigorii Zhislin and Aleksandr Sigalov

[105, 106] that for neutral molecules and positively charged ions, the electronic Hamiltonian possesses a lowest eigenvalue, frequently called the ground state energy, and a great deal of quantum chemical simulations are concerned with approximating this ground state energy.

The primary difficulty in the numerical computation of the lowest eigenvalue of the electronic Hamiltonian is the extremely high-dimensionality of the underlying Hilbert space. Indeed, for a system containing Q electrons, the sought-after ground state of the electronic Hamiltonian depends on $3Q$ spatial variables. A naive application of traditional numerical methods such as finite element approximations or spectral schemes, etc., therefore fails spectacularly, and specialised approximation strategies have to be developed. Broadly speaking, ab initio (first principles-based) deterministic numerical methods for approximating the ground state energy can be divided into three categories, each of which has a vast variety of subcategories and flavours (see, e.g., [15, Chapter 1] for a concise but comprehensive overview).

- Wave-function methods which focus on approximating directly the ground state of the electronic Hamiltonian.
- Density functional methods which are based on a reformulation of the minimisation problem for the electronic Hamiltonian (which acts on functions of $3Q$ spatial variables) in terms of an equivalent minimisation problem over a set of *electronic densities* (which are functions of 3 spatial variables).
- Reduced density matrix approaches which are based on the electronic one-body and two-body reduced density matrices.

The coupled cluster (CC) methodology, which belongs to the class of wave-function methods, is based on a non-linear ansatz for the sought-after ground state of the electronic Hamiltonian. In its most common form– the so-called single reference CC method– the unknown ground state is expressed as the action of an *exponential* cluster operator, i.e., the operator exponential of a linear combination of linear maps (so-called *excitation operators*), acting on a judiciously chosen reference function (usually a so-called discrete Hartree-Fock determinant). Using this ansatz the eigenvalue problem for the ground state energy of the electronic Hamiltonian can be reformulated as a *non-linear* system of equations for the unknown coefficients appearing in the linear combination of excitation operators entering the operator exponential. Approximations to the ground state energy are then obtained by restricting the class of excitation operators that appear inside the exponential, which leads to a hierarchy of computationally more tractable non-linear, root-finding problems. Usually these truncations are done on the basis of the excitation orders (see Chapter 2.2.3 below) and one thus speaks of CCD (double excitation operators only), CCSD (single and double excitation operators), CCSDT (single, double and triple excitation operators) and so on.

An important yet subtle concept in these post-Hartree Fock methods (including coupled cluster methods) deals with what is called "size consistency" (see e.g. [93]) and/or "size extensivity" see e.g. [52]). Intuitively, these properties are meant to describe whether or not the numerical method under consideration correctly captures the correct scaling properties of the system. Size-consistency concerns the additive separability of the energy of a molecular system, i.e., if we divide a molecular system into two non-interacting subsystems (e.g. separated by an infinite distance), then the energy of the whole system computed by a size-consistent method is the sum of the energies of these two subsystems taken separately. A size-extensive method, on the other hand, yields energies that scale linearly with the number of electrons. While these concepts may appear similar, they are not the same in general, and we refer to [70] for a careful presentation of these concepts starting from their historical introduction.

Focusing on the notion of size-consistency, it is well-known that without the exponential structure of the ansatz wave function, this property cannot be satisfied. On the contrary, as is proven by Reinhold Schneider in [87], the method we consider in this chapter (SRCC) is size-consistent under the assumption that the reference determinant is multiplicatively separable (see e.g., the work of Andreas Savin in [86] for the meaning of this term). Note that, in the vast literature, other couple cluster methods are said to benefit from size-consistency and/or size-extensivity even though proofs are missing and no explicit assumption is made about multiplicative separability of the reference determinant.

Coupled cluster methods were originally introduced in the field of nuclear physics in the late 1950s by Fritz Coester and Hermann Kummel [27, 28] but were reformulated for use in quantum chemistry in the following decades by pioneers such as Jiří Čížek [25], Josef Paldus [74], and Oktay Sinanoğlu [89]. The original motivation for introducing such methods was the fact that they satisfied the size-consistency property under the previously mentioned assumption regarding the reference determinant. Since size consistency seems to be a vital chemical property not conserved by some other numerical methods, and in practice, the CC methods seem to work extremely well, achieving, in many cases, the chemical accuracy of 1 kcal/mol, they quickly found wide adoption in the quantum chemical community [62]. In particular, the so-called CCSD(T)[1] variant, which can be applied to small and medium-sized molecules at a reasonable computational cost, is widely regarded as the 'gold standard' of quantum chemistry [79].

Despite the ubiquitous use of this 'gold standard' computational method in the quantum chemical community, there is a shockingly limited amount of mathematical literature on the numerical analysis of the coupled cluster methodology. Indeed, a simple search with the keyword "coupled cluster" on Google Scholar and MathSciNet, two databases that are representative of the scientific literature as a whole and the subset of mathematical literature thereof, reveals that there are more than 100,000 articles pertaining to coupled cluster theory of which less than 40 are listed on MathSciNet. Limiting ourselves to the subset of numerical analysis journals, there are a total seven articles on coupled cluster methods.

The first systematic study of the single reference coupled cluster method from a numerical analysis perspective was undertaken by Reinhold Schneider and Thorsten Rohwedder slightly more than ten years ago. In a series of three remarkable papers [87, 82, 83], they were able to show that the excitation operators that appear inside the coupled cluster exponential are bounded linear maps between Hilbert spaces of antisymmetric functions with appropriate regularity and that consequently, the continuous (infinite-dimensional) coupled cluster equations could be given a precise functional-analytic meaning. The coupled cluster approximations (built by restricting the class of excitation operators that enter the exponential operator) could thus be viewed as classical Galerkin discretization of an infinite-dimensional non-linear problem. Schneider also showed that under some assumptions, the underlying non-linear coupled cluster function is *locally, strongly monotone* which could exploited to prove the local well-posedness of both the continuous CC equations and its Galerkin discretisations. Schneider and Rohwedder also derived optimal error estimates for the coupled cluster energies using the dual-weighted residual approach of Rolf Rannacher and co-workers [92]. Since this pioneering work, two further contributions have been published which provide a similar numerical analysis for two other flavours of coupled cluster methods, namely, the *extended* coupled cluster method [61] and the *tailored* coupled cluster method [42] (see also [60]). In addition to the aforementioned contributions which tackle the coupled cluster equations from a functional analysis perspective, there has been recent interest in analysing the CC equations using tools from other fields. Thus, the contributions [29, 30] use concepts from graph theory to present a unified framework for constructing different variants

[1]Here, the (T) emphasises the fact that triple excitation orders are not initially included in the CCSD(T) ansatz and are rather treated perturbatively through a post-processing step.

of coupled cluster methods and topological index theory to study the solutions of the coupled cluster equations in finite-dimensions. Recently, an additional contribution has appeared which investigates the root structure of the CC equations using tools from algebraic geometry [43].

While the articles [87, 82, 83, 61, 42] listed above lay the groundwork for a rigorous *a priori* error analysis of the coupled cluster methods, they have one rather unfortunate drawback: in all cases, the well-posedness of the CC equations is established by demonstrating that the underlying CC function is locally strongly monotone, and this demonstration can only be shown to hold if the targeted root $\boldsymbol{t}^*$ of the CC function is sufficiently close to zero. In other words, the local well-posedness analysis and the resulting error estimates only hold in a perturbative regime $\boldsymbol{t}^* \approx 0$. On the other hand, as we discuss in more detail in Remark 2.4.2 in Chapter 2.4, in many practical situations where the CC method is *known numerically* to yield accurate approximations, the sought-after root $\boldsymbol{t}^*$ is *not* in the perturbative regime. For such problems, the existing *a priori* analysis yields estimates with *negative* constants! The *a priori* analysis having failed, there is also no hope of developing *a posteriori* error estimates for practical coupled cluster simulations which, in our opinion, would be the ultimate goal of the numerical analysis.

The aim of the current contribution is to develop a new *a priori* error analysis for the single reference coupled cluster equations that is valid under more general conditions. The analysis we present here– motivated by the existing literature on non-linear numerical analysis (see, for instance, [12, 96])– is based on the invertibility of the Fréchet derivative of the non-linear coupled cluster function, which is established using a classical inf-sup-type approach. In contrast to the local, strong monotonicity approach pioneered by Schneider, our analysis does not require the sought-after root $\boldsymbol{t}^*$ of the coupled cluster function to be close to zero. In our work, we will focus on the continuous (infinite-dimensional) CC equations and a specific version of the discrete CC equations, namely, the Full-CC equations in a finite basis (see Chapter 2.5). The extension of our analysis to more general discretization (the so-called *truncated* CC equations [52, Chapter 13]) will be addressed in a forthcoming contribution.

The remainder of this part is organized as follows. In Chapter 2.2, we introduce more rigorously the problem formulation, i.e., the electronic Hamiltonian and the Hilbert spaces on which it acts. In Chapter 2.3, we introduce excitation operators and the coupled cluster ansatz, and we state the continuous and discrete coupled cluster equations. We begin our analysis in Chapter 2.4 where we prove, under the minimal assumptions that the sought-after eigenfunction is intermediately normalizable and the associated eigenvalue is non-degenerate, that the continuous (infinite-dimensional) CC equations are *always* locally well-posed. In Chapter 2.5, we analyze a specific discretization of the CC equations, namely, the Full-CC equations in a finite basis. We prove under the same minimal assumptions of eigenpair non-degeneracy and CC ansatz validity that these equations are locally well-posed provided that the discretization is fine enough, and we derive residual-based error estimates with *guaranteed positive* constants. Preliminary numerical experiments indicate that the constants that appear in our estimates are a significant improvement over those obtained from the local monotonicity approach.

2.2 Problem formulation and setting

Computational quantum chemistry is the study of the properties of matter through modelling at the molecular scale, i.e., when matter is viewed as a collection of positively charged nuclei and negatively charged electrons. To formalise the problem setting, we assume that we are given a molecule composed of $M \in \mathbb{N}$ nuclei carrying charges $\{Z_n\}_{n=1}^M \subset \mathbb{R}_+$ and located at positions $\{\mathbf{z}_n\}_{n=1}^M \subset \mathbb{R}^3$, respectively. We further assume the presence of $Q \in \mathbb{N}$ electrons whose spatial

coordinates are denoted by $\{\mathbf{x}_i\}_{i=1}^{Q} \subset \mathbb{R}^3$. Throughout this thesis, we will assume that the Born-Oppenheimer approximation holds, i.e., we will treat the nuclei as fixed, classical particles and we will focus purely on the quantum mechanical description of the electrons.

In order to describe the behaviour of this system of nuclei and electrons under the Born-Oppenheimer approximation, we require the notion of several functions spaces. The following construction is partially based on [81].

2.2.1 Function spaces and norms

To begin with, we denote by $L^2(\mathbb{R}^3)$ the space of real-valued square integrable functions of three variables, and we denote by $H^1(\mathbb{R}^3)$ the closed subspace of $L^2(\mathbb{R}^3)$ consisting of functions that additionally possess square integrable first derivatives. Both spaces are equipped with their usual inner products. Following the convention in the quantum chemical literature, we will frequently refer to $L^2(\mathbb{R}^3)$ and $H^1(\mathbb{R}^3)$ as infinite-dimensional *single particle* spaces.

Next, we define the tensor space[2]

$$\mathcal{L}^2 := \bigotimes_{j=1}^{Q} L^2(\mathbb{R}^3),$$

which is equipped with an inner product that is constructed by defining first for all elementary tensors $f, g \in \mathcal{L}^2$ with $f = \otimes_{j=1}^{Q} f_j$ and $g = \otimes_{j=1}^{Q} g_j$

$$(f, g)_{\mathcal{L}^2} := \prod_{j=1}^{Q} (f_j, g_j)_{L^2(\mathbb{R}^3)}, \tag{2.1}$$

and then extending bilinearly for general tensorial elements of $\mathcal{L}^2$.

It is a consequence of Fubini's theorem that the tensor space $\mathcal{L}^2$ is isometrically isomorphic to the space $L^2(\mathbb{R}^{3Q})$ of real-valued square integrable functions of $3Q$ variables with the associated L^2-inner product. Thanks to this result, we can define the tensor space $\mathcal{H}^1 \subset \mathcal{L}^2$ as the closure of $\mathscr{C}_0^\infty(\mathbb{R}^{3Q})$ in $L^2(\mathbb{R}^{3Q})$ with respect to the usual gradient-gradient inner product on $\mathbb{R}^{3Q}$.

In quantum mechanics, a fundamental distinction is made between so-called *bosonic* and *fermionic* particles, the latter obeying the so-called Pauli-exclusion principle and thus being described in terms of antisymmetric functions. We are therefore obligated to also define tensor spaces of antisymmetric functions. To this end, we first introduce the so-called *antisymmetric projection operator* $\mathbb{P}^{\mathrm{as}} \colon \mathcal{L}^2 \to \mathcal{L}^2$ that is defined through the action

$$\forall f \in \mathcal{L}^2 \colon \quad (\mathbb{P}^{\mathrm{as}} f)(\mathbf{x}_1, \ldots, \mathbf{x}_Q) := \frac{1}{Q!} \sum_{\pi \in \mathrm{S}(Q)} (-1)^{\mathrm{sgn}(\pi)} f(\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(Q)}),$$

where $\mathrm{S}(Q)$ denotes the permutation group of order Q, and $\mathrm{sgn}(\pi)$ denotes the signature of $\pi \in \mathrm{S}(Q)$.

[2] We remind the reader that topological tensor spaces can be defined by taking an orthonormal basis of the underlying single particle spaces (in this case $L^2(\mathbb{R}^3)$) and using it to construct the algebraic tensor product vector space. This vector space is equipped with a tensorial inner product inherited from the single particle function spaces (in this case the inner product is given by Equation (2.1)). The topological tensor product space is then obtained by taking the completion of the algebraic tensor product vector space with respect to the norm induced by the tensorial inner product.

It is easy to establish that $\mathbb{P}^{\text{as}}$ is an $\mathcal{L}^2$-orthogonal projection with a closed range. We therefore define the antisymmetric tensor spaces $\widehat{\mathcal{L}^2} \subset \mathcal{L}^2$ and $\widehat{\mathcal{H}^1} \subset \mathcal{H}^1$ as

$$\widehat{\mathcal{L}^2} := \bigwedge_{j=1}^{Q} L^2(\mathbb{R}^3) := \text{ran } \mathbb{P}^{\text{as}} \quad \text{and} \quad \widehat{\mathcal{H}^1} := \widehat{\mathcal{L}^2} \cap \mathcal{H}^1,$$

equipped with the $(\cdot,\cdot)_{\mathcal{L}^2}$ and $(\cdot,\cdot)_{\mathcal{H}^1}$ inner products respectively. We remark that normalised elements of $\widehat{\mathcal{L}^2}$ are known as *wave-functions*, and these are antisymmetric in the sense that for any $f \in \widehat{\mathcal{L}^2}$ we have that

$$f(\mathbf{x}_1, \ldots, \mathbf{x}_i, \ldots, \mathbf{x}_j, \ldots \mathbf{x}_Q) = -f(\mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_i, \ldots \mathbf{x}_Q) \qquad \forall\, i, j \in \{1, \ldots, Q\} \text{ with } i \neq j.$$

In the sequel, we will also occasionally make use of the dual space of $\widehat{\mathcal{H}^1}$. We therefore denote $\widehat{\mathcal{H}^{-1}} := \big(\widehat{\mathcal{H}^1}\big)^*$, we equip $\widehat{\mathcal{H}^{-1}}$ with the canonical dual norm, and we write $\langle \cdot, \cdot \rangle_{\widehat{\mathcal{H}^1}, \widehat{\mathcal{H}^{-1}}}$ for the associated duality pairing. Note that higher regularity Sobolev spaces $\widehat{\mathcal{H}^r}$, $r \geqslant 1$ can be defined similarly to $\widehat{\mathcal{H}^1}$.

Finally, let us comment on the construction of basis sets for the tensor spaces $\mathcal{H}^1$ and $\widehat{\mathcal{H}^1}$. Given an L^2-orthonormal, complete basis $\mathcal{B} := \{\phi_k\}_{k \in \mathbb{N}} \subset H^1(\mathbb{R}^3)$, we can construct a complete basis $\mathcal{B}_\otimes$ for $\mathcal{H}^1$ by setting

$$\mathcal{B}_\otimes = \Big\{ \phi_{k_1} \otimes \phi_{k_2} \otimes \ldots \otimes \phi_{k_Q} : \; k_1, k_2, \ldots, k_Q \in \mathbb{N} \Big\},$$

and it follows immediately that $\mathcal{B}_\otimes$ is $\mathcal{L}^2$-orthonormal.

In order to construct a basis for the antisymmetric tensor space $\widehat{\mathcal{H}^1}$, we must first define a suitable subset of $\mathcal{B}_\otimes$. To this end, we introduce an index set $\mathcal{J}_\infty^Q \subset \mathbb{N}^Q$ given by

$$\mathcal{J}_\infty^Q := \Big\{ \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_Q) \in \mathbb{N}^Q \colon \alpha_1 < \alpha_2 < \ldots < \alpha_Q \Big\}.$$

We can thus define the subset $\mathcal{B}_\otimes^{\text{ord}}$ of the basis $\mathcal{B}_\otimes$ given by

$$\mathcal{B}_\otimes^{\text{ord}} := \Big\{ \widetilde{\Phi_\alpha} := \phi_{\alpha_1} \otimes \phi_{\alpha_2} \otimes \ldots \otimes \phi_{\alpha_Q} \colon \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_Q) \in \mathcal{J}_\infty^Q \Big\}.$$

A complete basis for the antisymmetric tensor space $\widehat{\mathcal{H}^1}$ is then given by

$$\begin{aligned} \mathcal{B}_\wedge :=& \Big\{ \frac{\mathbb{P}^{\text{as}} \Phi}{\|\mathbb{P}^{\text{as}} \Phi\|_{\widehat{\mathcal{L}^2}}} : \Phi \in \mathcal{B}_\otimes^{\text{ord}} \Big\} \\ =& \left\{ \Phi_\alpha(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \sum_{\pi \in \mathrm{S}(Q)} (-1)^{\text{sgn}(\pi)} \otimes_{i=1}^{Q} \phi_{\alpha_i}\big(\mathbf{x}_{\pi(i)}\big) : \; \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_Q) \in \mathcal{J}_\infty^Q \right\}. \end{aligned}$$

Elements of the basis set $\mathcal{B}_\wedge$ are called Slater determinants. For simplicity, given $\alpha \in \mathcal{J}_\infty^Q$ and $\Phi_\alpha \in \mathcal{B}_\wedge$ of the form

$$\Phi_\alpha(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \sum_{\pi \in \mathrm{S}(Q)} (-1)^{\text{sgn}(\pi)} \otimes_{i=1}^{Q} \phi_{\alpha_i}\big(\mathbf{x}_{\pi(i)}\big),$$

we will write Φ_α in the succinct form

$$\Phi_\alpha(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \det\big(\phi_{\alpha_i}(\mathbf{x}_j)\big)_{i,j=1}^{Q}.$$

2.2.2 Governing operators and problem statement

Throughout this thesis, we assume that the electronic properties of the molecule that we study can be described by the action of a many-body electronic Hamiltonian given by

$$H := -\frac{1}{2}\sum_{j=1}^{Q} \Delta_{\mathbf{x}_j} + \sum_{j=1}^{Q}\sum_{n=1}^{M} \frac{-Z_n}{|\mathbf{z}_n - \mathbf{x}_j|} + \sum_{j=1}^{Q}\sum_{i=1}^{j-1} \frac{1}{|\mathbf{x}_i - \mathbf{x}_j|} \qquad \text{acting on } \widehat{\mathcal{L}}^2 \quad \text{with domain } \widehat{\mathcal{H}}^2. \tag{2.2}$$

The electronic properties of the molecule that we study are functions of the spectrum of the electronic Hamiltonian H, and we are therefore interested in its analysis and computation. It is a classical result (see, e.g., the review article [55]) that the operator H is self-adjoint on $\widehat{\mathcal{L}}^2$ with form domain $\widehat{\mathcal{H}}^1$, and under the additional assumption that $Z := \sum_{n=1}^{M} Z_n \geqslant Q$, it holds that

1. The operator H has an essential spectrum σ_{ess} of the form $\sigma_{\text{ess}} := [\Sigma, \infty)$ where $-\infty < \Sigma \leqslant 0$;

2. The operator H has a bounded-below discrete spectrum that consists of a countably infinite number of eigenvalues, each with finite multiplicity, accumulating at Σ.

Consequently, under the assumption that $\sum_{n=1}^{M} Z_n \geqslant Q$, the electronic Hamiltonian $\mathcal{H}$ possesses a lowest eigenvalue $\mathcal{E}^*_{\text{GS}} \in \mathbb{R}$, frequently called the ground state energy, such that

$$\mathcal{E}^*_{\text{GS}} = \min_{0 \neq \Psi \in \widehat{\mathcal{H}}^1} \frac{\big\langle \Psi, H\Psi \big\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}}{\|\Psi\|^2_{\mathcal{L}^2}}. \tag{2.3a}$$

Any function $\Psi^*_{\text{GS}} \in \widehat{\mathcal{H}}^1$ that achieves the minimum in Equation (2.3a) is called a ground state of H and obviously satisfies

$$H\Psi^*_{\text{GS}} = \mathcal{E}^*_{\text{GS}}\Psi^*_{\text{GS}}. \tag{2.3b}$$

For the purpose of this thesis, we will assume that indeed $Z := \sum_{n=1}^{M} Z_n \geqslant Q$. Note that if the ground state eigenvalue $\mathcal{E}^*_{\text{GS}}$ is simple (which is not always the case), normalised ground states Ψ^*_{GS} (being elements of a real Hilbert space) are unique up to sign.

From a functional analysis point of view, the electronic Hamiltonian H possesses certain desirable properties, namely continuity and ellipticity on appropriate Sobolev spaces. More precisely (see, for instance, [103, Chapter 4]),

- The electronic Hamiltonian defined through Equation (2.2) is bounded as a mapping from $\widehat{\mathcal{H}}^1$ to $\widehat{\mathcal{H}}^{-1}$:

$$\forall \Phi, \Psi \in \widehat{\mathcal{H}}^1: \qquad \Big|\big\langle \Phi, H\Psi \big\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}\Big| \leqslant \Big(\frac{1}{2} + 3\sqrt{Q}Z\Big)\|\Phi\|_{\widehat{\mathcal{H}}^1}\|\Psi\|_{\widehat{\mathcal{H}}^1}; \tag{2.4}$$

- The electronic Hamiltonian defined through Equation (2.2) satisfies the following ellipticity condition on the Gelfand triple $\widehat{\mathcal{H}}^1 \hookrightarrow \widehat{\mathcal{L}}^2 \hookrightarrow \widehat{\mathcal{H}}^{-1}$:

$$\forall \Phi \in \widehat{\mathcal{H}}^1 : \qquad \langle \Phi, H\Phi \rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} \geqslant \frac{1}{4} \|\Phi\|^2_{\widehat{\mathcal{H}}^1} - \Big(9QZ^2 - \frac{1}{4}\Big) \|\Phi\|^2_{\widehat{\mathcal{L}}^2}. \tag{2.5}$$

An important consequence of the above ellipticity estimate is that the electronic Hamiltonian, modified by any suitable shift, defines an invertible operator on a subspace of $\widehat{\mathcal{H}}^1$. This fact will be of great importance in our analysis and will be the subject of further discussion in Chapter 2.4 (see Remark 2.4.3).

Remark 2.2.1 (Restriction to function spaces of real-valued, spin-independent functions)**.**
To avoid notational complexity, we have restricted our analysis in this thesis to real-valued function spaces and we have not taken into account spin variables. The restriction to real-valued functions does not result in any loss in generality since the governing operator that we consider, namely the electronic Hamiltonian H defined through Equation (2.2)*, consists entirely of real terms, and thus the real and imaginary parts of any eigenfunction of H are themselves eigenfunctions of H (see, e.g., [15, Chapter 6.1] for a brief discussion of this point).*

Our choice to neglect spin is motivated by the same observation, i.e., that the electronic Hamiltonian does not contain any explicit spin dependencies. Consequently, in order to take spin variables into account we need simply replace the single particle function spaces

$$L^2(\mathbb{R}^3) \;\; \textit{and} \;\; H^1(\mathbb{R}^3) \qquad \textit{with} \qquad L^2\Big(\mathbb{R}^3 \times \Big\{ \pm \frac{1}{2} \Big\}\Big) \;\; \textit{and} \;\; H^1\Big(\mathbb{R}^3 \times \Big\{ \pm \frac{1}{2} \Big\}\Big) \;\; \textit{respectively.}$$

Here, $L^2\left(\mathbb{R}^3 \times \left\{\pm\frac{1}{2}\right\}\right)$ can be seen as the space of (equivalence classes of) square-integrable functions of three spatial variables and an additional spin variable $s = \pm\frac{1}{2}$, equipped with the inner product

$$\forall f, g \in L^2\Big(\mathbb{R}^3 \times \Big\{ \pm \frac{1}{2} \Big\}\Big) : \quad (f,g)_{L^2\left(\mathbb{R}^3 \times \left\{\pm\frac{1}{2}\right\}\right)} = \sum_{s=\pm\frac{1}{2}} \int_{\mathbf{R}^3} f(\mathbf{x}, s) g(\mathbf{x}, s) \, d\mathbf{x},$$

and an analogous interpretation holds for $H^1\left(\mathbb{R}^3 \times \left\{\pm\frac{1}{2}\right\}\right)$.

Equipped with the spin-dependent single particle spaces $L^2\left(\mathbb{R}^3 \times \left\{\pm\frac{1}{2}\right\}\right)$ and $H^1\left(\mathbb{R}^3 \times \left\{\pm\frac{1}{2}\right\}\right)$, the spin-dependent tensorial Q-particle function spaces and basis sets can be constructed following mutatis mutandis, the procedure described in Chapter 2.2.1 above. Our subsequent analysis can then be readily applied to such spin-dependent function spaces without any significant modifications (with only slight modification of the values of some constants appearing in our following error estimates).

One additional feature of the spin-dependent formalism deserves mention. The analysis that we present in this contribution frequently requires assumptions on the simplicity of certain eigenvalues of the electronic Hamiltonian H. Unfortunately, considering H as an operator on the full spin-dependent tensorial space $\widehat{\mathcal{L}}^2_{\mathrm{s}} = \wedge_{j=1}^{Q} L^2\left(\mathbb{R}^3 \times \left\{\pm\frac{1}{2}\right\}\right)$ often introduces degeneracies in these eigenvalues, and in order to remove these degeneracies, it is necessary to restrict the funtional setting to a suitable subspace of $\widehat{\mathcal{L}}^2_{\mathrm{s}}$, typically an eigenspace of the so-called z-spin operator (see [81] for a detailed construction).

2.2.3 Computing the ground state energy in a finite-dimensional subspace

From a practical point of view, the ground state energy of the electronic Hamiltonian defined through Equation (2.2) can only be approximated in a finite-dimensional subspace. The most conceptually simple such approach (albeit tremendously computationally expensive and therefore not widely used) is known in the quantum chemical literature as *Full Configuration Interaction*. In this subsection, we introduce the terminology and briefly discuss the methodology of the full configuration interaction procedure since the underlying notions will be useful when we discuss the discrete coupled cluster equations in Chapter 2.5

At its core, the full configuration interaction method (Full-CI) is based on a straightforward Galerkin approximation of the minimisation problem (2.3a). We will therefore begin by defining an approximation space. To do so, we fix some $K \in \mathbb{N}$ with $K > Q$ and assume that we are given a set $\{\phi_j\}_{j=1}^K \subset H^1(\mathbb{R}^3)$ of $L^2(\mathbb{R}^3)$-orthonormal functions. We also introduce an index set $\mathcal{I}_K^Q \subset \{1,\dots,K\}^Q$ given by

$$\mathcal{I}_K^Q := \Big\{\alpha = (\alpha_1, \alpha_2, \dots, \alpha_Q) \in \{1,\dots,K\}^Q \colon \alpha_1 < \alpha_2 < \dots < \alpha_Q\Big\}.$$

Definition 2.2.1 (Finite Dimensional Single-Particle Basis)**.**
We define the K-dimensional single particle basis $\mathcal{B}_K \subset H^1(\mathbb{R}^3)$ as $\mathcal{B}_K := \{\phi_j\}_{j=1}^K$. Additionally, we define the subspace spanned by this basis set as $X_K := \operatorname{span} \mathcal{B}_K$ and we refer to X_K as the single particle approximation space.

Definition 2.2.2 (Finite Dimensional Q-Particle Basis)**.**
We define the $\mathcal{L}^2$-orthonormal, $\binom{K}{Q}$-dimensional Q-particle basis $\mathcal{B}_K^Q \subset \widehat{\mathcal{H}}^1$ as

$$\mathcal{B}_K^Q := \left\{\Phi_\alpha(\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \det\big(\phi_{\alpha_i}(\mathbf{x}_j)\big)_{i,j=1}^Q \colon\ \alpha = (\alpha_1, \alpha_2, \dots, \alpha_Q) \in \mathcal{I}_K^Q\right\}.$$

Additionally, we define the subspace spanned by this basis set as $\mathcal{V}_K := \operatorname{span} \mathcal{B}_K^Q$ and we refer to $\mathcal{V}_K$ as the Q-particle approximation space.

Full Configuration Interaction Approximation of Minimisation Problem (2.3a)
Let the Q-particle approximation space $\mathcal{V}_K$ be defined through Definition 2.2.2. We seek the pair(s) $(\mathcal{E}_{\mathrm{FCI}}^*, \Psi_{\mathrm{FCI}}^*) \in \big(\mathbb{R}, \mathcal{V}_K\big)$ with $\|\Psi_{\mathrm{FCI}}^*\|_{\mathcal{L}^2}^2 = 1$ that satisfies

$$\mathcal{E}_{\mathrm{FCI}}^* := \min_{0 \neq \Psi \in \mathcal{V}_K} \frac{\langle \Psi, H\Psi\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}}{\|\Psi\|_{\mathcal{L}^2}^2} \quad \text{and} \quad \langle \Psi, H\Psi_{\mathrm{FCI}}^*\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = \mathcal{E}_{\mathrm{FCI}}^* \left(\Psi, \Psi_{\mathrm{FCI}}^*\right)_{\widehat{\mathcal{L}}^2} \quad \forall \Psi \in \mathcal{V}_K. \tag{2.6}$$

Several remarks are now in order.

First, it follows from the variational principle that the minimum in Equation (2.6) satisfies $\mathcal{E}_{\mathrm{FCI}}^* \geqslant \mathcal{E}_{\mathrm{GS}}^*$.

Second, in practice the Full-CI minimisation problem (2.6) is very often solved by writing first the associated Euler-Lagrange equations, i.e., the first order optimality conditions. This yields a linear eigenvalue problem on the finite-dimensional space $\mathcal{V}_K$ which can, in principle, be solved through the use of some iterative eigenvalue solver.

Third, despite the fact that the Full-CI methodology (2.6) seems very amenable to numerical analysis- by virtue of being a Galerkin approximation to the exact minimisation problem (2.3a)- it has a fundamental computational draw-back: the dimension of the Q-particle approximation space $\mathcal{V}_K$ grows combinatorially in Q which renders this approach computationally intractable for Q even moderately large. As a consequence, we are very often forced to introduce further approximations to the Full-CI methodology.

We end this section by defining the Full-CI Hamiltonian which will be referenced in Chapter 2.5 below.

Definition 2.2.3 (Full-CI Hamiltonian)**.**
Let the Q-particle approximation space $\mathcal{V}_K$ be defined through Definition 2.2.2 and let the electronic Hamiltonian $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ be defined through Equation (2.2). We define the Full-CI Hamiltonian $H_K\colon \mathcal{V}_K \to \mathcal{V}_K^*$ as the mapping with the property that for all $\Psi_K, \Phi_K \in \mathcal{V}_K$ it holds that

$$\langle \Psi_K, H_K \Phi_K \rangle_{\mathcal{V}_K \times \mathcal{V}_K^*} := \langle \Psi_K, H\Phi_K \rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}. \tag{2.7}$$

2.3 Excitation operators and the coupled cluster ansatz

Throughout this section, we assume the setting of Chapter 2.2. Our goal now is to introduce the notions of excitation indices, excitation operators and the coupled cluster non-linear parameterisation ansatz.

Let us begin by recalling that we have introduced *complete* single-particle and Q-particle basis sets $\mathcal{B} = \{\phi_j\}_{j\in\mathbb{N}} \subset H^1(\mathbb{R}^3)$ and $\mathcal{B}_\wedge \subset \widehat{\mathcal{H}}^1$ respectively in Chapter 2.2.1. Next, we define a collection of index sets.

Definition 2.3.1 (Excitation Index Sets)**.**
For each $j \in \{1, \dots, Q\}$ we define the index set $\mathcal{I}_j$ as

$$\mathcal{I}_j := \left\{ \binom{a_1, \dots, a_j}{\ell_1, \dots, \ell_j} \colon \ell_1 < \dots < \ell_j \in \{1, \dots, Q\} \text{ and } a_1 < \dots < a_j \in \{Q+1, Q+2, \dots\} \right\},$$

and we say that $\mathcal{I}_j$ is the excitation index set of order j. Additionally, we define

$$\mathcal{I} := \bigcup_{j=1}^{Q} \mathcal{I}_j,$$

and we say that $\mathcal{I}$ is the global excitation index set.

The excitation index sets $\{\mathcal{I}_j\}_{j=1}^Q$ will be used to construct the so-called excitation and de-excitation operators which play a central role in post-Hartree Fock wave-function methods for further approximating the minimisation problem (2.6).

Definition 2.3.2 (Excitation Operators)**.**
Let $j \in \mathbb{N}$ and let $\mu \in \mathcal{I}_j$ be of the form

$$\mu = \binom{a_1, \dots, a_j}{\ell_1, \dots, \ell_j} \colon \ell_1 < \dots < \ell_j \in \{1, \dots, Q\} \text{ and } a_1 < \dots < a_j \in \{Q+1, Q+2, \dots\}.$$

We define the excitation operator $\mathcal{X}_\mu\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ through its action on the Q-particle basis set $\mathcal{B}_\wedge$: For $\Psi_\nu(\mathbf{x}_1, \dots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \det\left(\phi_{\nu_j}(\mathbf{x}_i)\right)_{i,j=1}^Q$, we set

$$\mathcal{X}_\mu \Psi_\nu = \begin{cases} 0 & \text{if } \{\ell_1, \dots, \ell_j\} \not\subset \{\nu_1, \dots, \nu_Q\}, \\ 0 & \text{if } \exists a_m \in \{a_1, \dots, a_j\} \text{ such that } a_m \in \{\nu_1, \dots, \nu_Q\}, \\ \Psi_\nu^a & \text{otherwise,} \end{cases}$$

where the determinant Ψ_ν^a is constructed from Ψ_ν by replacing all functions $\phi_{\ell_1}, \dots \phi_{\ell_j}$ used to construct Ψ_ν with functions $\phi_{a_1}, \dots, \phi_{a_j}$ respectively. In addition, Ψ_ν^a is an element of the basis set $\mathcal{B}_\wedge$ up to a sign depending on its index ordering.

Definition 2.3.3 (De-excitation Operators)**.**
Let $j \in \mathbb{N}$ and let $\mu \in \mathcal{I}_j$ be of the form

$$\mu = \binom{a_1, \dots, a_j}{\ell_1, \dots, \ell_j} \colon \ell_1 < \dots < \ell_j \in \{1, \dots, Q\} \text{ and } a_1 < \dots < a_j \in \{Q+1, Q+2, \dots\}.$$

We define the de-excitation operator $\mathcal{X}_\mu^\dagger \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ through its action on the Q-particle basis set $\mathcal{B}_\wedge$: For $\Psi_\nu(\mathbf{x}_1, \dots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \det\left(\phi_{\nu_j}(\mathbf{x}_i)\right)_{i,j=1}^Q$, we set

$$\mathcal{X}_\mu^\dagger \Psi_\nu = \begin{cases} 0 & \text{if } \{a_1, \dots, a_j\} \not\subset \{\nu_1, \dots, \nu_Q\}, \\ 0 & \text{if } \exists\ \ell_m \in \{\ell_1, \dots, \ell_j\} \text{ such that } \ell_m \in \{\nu_1, \dots, \nu_Q\}, \\ \Psi_{\nu,\ell} & \text{otherwise,} \end{cases}$$

where the determinant $\Psi_{\nu,\ell}$ is constructed from Ψ_ν by replacing all functions $\phi_{a_1}, \dots \phi_{a_j}$ used to construct Ψ_ν with functions $\phi_{\ell_1}, \dots, \phi_{\ell_j}$ respectively. In addition, $\Psi_{\nu,\ell}$ is an element of the basis set $\mathcal{B}_\wedge$ up to a sign depending on its index ordering.

It is natural to ask how de-excitation operators are related to excitation operators. The following remark summarises this relationship.

Remark 2.3.1 (Relationship between Excitation and De-excitation Operators)**.**
Consider the setting of Definitions 2.3.2 and 2.3.3. In some sense, each de-excitation operator reverses the action the corresponding excitation operator. More precisely, it can be shown that for any $\mu \in \mathcal{I}$, the de-excitation operator $\mathcal{X}_\mu^\dagger \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ is the $\widehat{\mathcal{L}}^2$-adjoint of the excitation operator $\mathcal{X}_\mu \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$, i.e.,

$$\forall \Phi, \widetilde{\Phi} \in \widehat{\mathcal{H}}^1,\ \forall \mu \in \mathcal{I} \colon \qquad \left(\widetilde{\Phi}, \mathcal{X}_\mu \Phi\right)_{\widehat{\mathcal{L}}^2} = \left(\mathcal{X}_\mu^\dagger \widetilde{\Phi}, \Phi\right)_{\widehat{\mathcal{L}}^2}.$$

Several properties of the excitation operators can now be deduced. We begin with a remark.

Remark 2.3.2 (Interpretation of Q-particle Basis in Terms of Excited Determinants)**.**
It is a simple exercise to show that any Slater determinant in the Q-particle basis set $\mathcal{B}_\wedge$ can be generated through the action of the excitation operators on a so-called reference determinant up to a sign depending on its index ordering. More precisely, we define $\Psi_0(\mathbf{x}_1, \dots, \mathbf{x}_Q) := \frac{1}{\sqrt{Q!}} \det\left(\phi_j(\mathbf{x}_i)\right)_{i,j=1}^Q$, and it then follows that

$$\begin{aligned} \mathcal{B}_\wedge &= \{\Psi_0\} \cup \{\mathcal{X}_\mu \Psi_0 \colon \mu \in \mathcal{I}_1\} \cup \{\mathcal{X}_\mu \Psi_0 \colon \mu \in \mathcal{I}_2\} \cup \dots \cup \{\mathcal{X}_\mu \Psi_0 \colon \mu \in \mathcal{I}_Q\} \\ &= \{\Psi_0\} \cup \{\mathcal{X}_\mu \Psi_0 \colon \mu \in \mathcal{I}\}, \end{aligned} \tag{2.8}$$

up to a sign depending on the index ordering of the generated determinant. This observation motivates the following convention and definition.

Convention 2.3.1 (Reference Determinant)**.**
Consider the setting of Remark 2.3.2. In the sequel, we will refer to the function $\mathcal{B}_\wedge \ni \Psi_0(\mathbf{x}_1, \dots, \mathbf{x}_Q)$
$= \frac{1}{\sqrt{Q!}} \det\left(\phi_j(\mathbf{x}_i)\right)_{i,j=1}^Q$, i.e., the determinant constructed from the first Q single particle basis functions $\{\phi_i\}_{i=1}^Q$ as the reference determinant. Moreover, for any $\mu \in \mathcal{I}$, we will frequently

denote $\Psi_\mu := \mathcal{X}_\mu \Psi_0$. *Finally, we will often refer to each set* $\{\mathcal{X}_\mu \Psi_0 \colon \mu \in \mathcal{I}_j\}$ *as the set of* j*-excited determinants,*

Definition 2.3.4 (Orthogonal Complement of the Reference Determinant)**.**

Let the excitation index set $\mathcal{I}$ *be defined through Definition 2.3.1, let the excitation operators* $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{I}}$ *be defined through Definition 2.3.2, and let* $\Psi_0(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) := \frac{1}{\sqrt{Q!}} \det\left(\phi_j(\mathbf{x}_i)\right)_{i,j=1}^{Q}$ *denote the reference determinant. Then we define the set* $\widetilde{\mathcal{B}_\wedge} \subset \mathcal{B}_\wedge$ *and the subspace* $\widetilde{\mathcal{V}} \subset \widehat{\mathcal{H}}^1$ *as*

$$\begin{aligned} \widetilde{\mathcal{B}_\wedge} &:= \{\mathcal{X}_\mu \Psi_0 \colon \mu \in \mathcal{I}\}, \qquad \text{and} \\ \widetilde{\mathcal{V}} &:= \{\Psi_0\}^\perp := \left\{\Phi \in \widehat{\mathcal{H}}^1 \colon \quad (\Phi, \Psi_0)_{\widehat{\mathcal{L}}^2} = 0\right\}, \end{aligned}$$

and we observe that $\widetilde{\mathcal{B}_\wedge}$ *is a complete,* $\widehat{\mathcal{L}}^2$*-orthonormal basis for* $\widetilde{\mathcal{V}}$.

Definition 2.3.5 (Complementary Decomposition of $\widehat{\mathcal{H}}^1$)**.**

Let the excitation index set $\mathcal{I}$ *be defined through Definition 2.3.1, let the excitation operators* $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{I}}$ *be defined through Definition 2.3.2, and let* $\Psi_0(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) := \frac{1}{\sqrt{Q!}} \det\left(\phi_j(\mathbf{x}_i)\right)_{i,j=1}^{Q}$ *denote the reference determinant. We define* $\mathbb{P}_0 \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ *as the* $\widehat{\mathcal{L}}^2$*-orthogonal projection operator onto span*$\{\Psi_0\}$*, and we define* $\mathbb{P}_0^\perp := \mathbb{I} - \mathbb{P}_0 \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ *as its complement. Additionally, we introduce the complementary decomposition of the* Q*-particle space* $\widehat{\mathcal{H}}^1$ *given by*

$$\widehat{\mathcal{H}}^1 = \operatorname{span}\{\Psi_0\} \oplus \widetilde{\mathcal{V}}, \quad \text{where we emphasise that } \widetilde{\mathcal{V}} = \operatorname{Ran}\mathbb{P}_0^\perp. \tag{2.9}$$

The complementary decomposition introduced through Equation (2.9) will be particularly important in our subsequent analysis of the coupled cluster method in Chapter 2.4. Let us emphasise that the construction of these complementary spaces is based on $\widehat{\mathcal{L}}^2$-orthogonality rather than $\widehat{\mathcal{H}}^1$ orthogonality. This choice is intentional as it simplifies considerably the analysis in Chapters 2.4 and 2.5. Let us also remark that the projection operators $\mathbb{P}_0 \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ and $\mathbb{P}_0^\perp \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ are both, nevertheless, bounded operators with respect to the $\|\cdot\|_{\widehat{\mathcal{H}}^1}$ norm since they both possess a closed range and a closed kernel.

Returning for the moment to the notion of excitation operators, we see that it is easy to deduce that each excitation operator $\mathcal{X}_\mu$, $\mu \in \mathcal{I}$ is a bounded linear operator from $\widehat{\mathcal{H}}^1$ to $\widehat{\mathcal{H}}^1$. However, we will frequently be interested in so-called cluster operators which are summations of the excitation operators $\mathcal{X}_\mu$, $\mu \in \mathcal{I}$, and such summations need not be bounded operators from $\widehat{\mathcal{H}}^1$ to $\widehat{\mathcal{H}}^1$ or even from $\widehat{\mathcal{L}}^2$ to $\widehat{\mathcal{L}}^2$. Fortunately, the following result was proven in [82].

Proposition 2.3.1 (Cluster Operators as Bounded Maps on $\widehat{\mathcal{L}}^2$)**.**

Let the excitation index set $\mathcal{I}$ *be defined through Definition 2.3.1 and let* $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \ell^2(\mathcal{I})$. *Then there exists a unique bounded linear operator* $\mathcal{T} \colon \widehat{\mathcal{L}}^2 \to \widehat{\mathcal{L}}^2$*, the so-called* cluster operator *generated by* $\boldsymbol{t}$*, such that* $\mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu$ *where the series convergence holds with respect to the operator norm* $\|\cdot\|_{\widehat{\mathcal{L}}^2 \to \widehat{\mathcal{L}}^2}$.

Next, we introduce a coefficient subspace of $\ell^2(\mathcal{I})$, i.e, the space of square summable sequences of real numbers indexed by $\mathcal{I}$, which will limit the class of cluster operators that we consider in the sequel.

Definition 2.3.6 (Coefficient Space For Cluster Operators)**.**

Let the excitation index set $\mathcal{I}$ *be defined through Definition 2.3.1 and let* $\ell^2(\mathcal{I})$ *denote the space of square summable sequences of real numbers indexed by* $\mathcal{I}$. *We define the Hilbert space*

of sequences $\mathbb{V} \subset \ell^2(\mathcal{I})$ *as the set*

$$\mathbb{V} := \left\{ \mathbf{t} := \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \ell^2(\mathcal{I}) : \quad \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \Psi_\mu \ \in \widehat{\mathcal{H}}^1 \right\}, \tag{2.10}$$

equipped with the inner product

$$\forall \mathbf{t}, \boldsymbol{s} \in \mathbb{V} \text{ with } \ \boldsymbol{t} := (\boldsymbol{t}_\mu)_{\mu\in\mathcal{I}}, \ \boldsymbol{s} := (\boldsymbol{s}_\nu)_{\nu\in\mathcal{I}} : \qquad (\boldsymbol{s},\boldsymbol{t})_{\mathbb{V}} := \left(\sum_{\mu\in\mathcal{I}} \boldsymbol{s}_\mu \Psi_\mu, \sum_{\nu\in\mathcal{I}} \boldsymbol{t}_\nu \Psi_\nu \right)_{\widehat{\mathcal{H}}^1}. \tag{2.11}$$

Additionally, we define $\mathbb{V}^*$ *as the topological dual space of* $\mathbb{V}$*, equipped with the canonical dual norm*

$$\forall \boldsymbol{w} \in \mathbb{V}^* : \qquad \|\boldsymbol{w}\|_{\mathbb{V}^*} := \sup_{0\neq \boldsymbol{t}\in\mathbb{V}} \frac{\left| \langle \boldsymbol{w}, \boldsymbol{t} \rangle_{\mathbb{V}^*\times\mathbb{V}} \right|}{\|\boldsymbol{t}\|_{\mathbb{V}}},$$

Some remarks are now in order.

Remark 2.3.3 (Clarification of the Definition of the Coefficient Space)**.**

Consider Definition 2.3.6 of the coefficient space $\mathbb{V} \subset \ell^2(\mathcal{I})$ *and let* $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$*. We emphasise here that the assertion* $\sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \Psi_\mu \in \widehat{\mathcal{H}}^1$ *should be understood in the following sense: there exists* $\Psi_{\boldsymbol{t}} \in \widehat{\mathcal{H}}^1 \subset \widehat{\mathcal{L}}^2$ *such that* $\Psi_{\boldsymbol{t}} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \Psi_\mu$ *where the series convergence holds with respect to the* $\widehat{\mathcal{L}}^2$ *norm. In particular, this series convergence does not* a priori *hold with respect to the* $\widehat{\mathcal{H}}^1$*-norm, and it is only the limit function* $\Psi_{\boldsymbol{t}}$ *that is an element of* $\widehat{\mathcal{H}}^1$*.*

Remark 2.3.4 (Dual Coefficient Space)**.**

Consider the setting of Definition 2.3.6. Throughout, this thesis, we will denote by $\mathbb{V}^*$ *the topological dual space of* $\mathbb{V}$ *equipped with the canonical dual norm. Note that since* $\widehat{\mathcal{H}}^1$ *is dense and continuously embedded in* $\widehat{\mathcal{L}}^2$*, we can deduce that the coefficient space* $\mathbb{V}$ *is dense and continuously embedded in* $\ell^2(\mathcal{I})$*. As a consequence, the inner product* $(\cdot,\cdot)_{\ell^2}$ *can be continuously extended to the duality pairing* $\langle \cdot,\cdot \rangle_{\mathbb{V}\times\mathbb{V}^*}$ *on* $\mathbb{V}\times\mathbb{V}^*$*. This fact will be of occasional use in the sequel.*

Notation 2.3.2 (Coefficient Sequences and Cluster Operators)**.**

Let $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$*. As mentioned previously, the operator* $\mathcal{T} := \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu$ *is known as the cluster operator generated by* $\boldsymbol{t}$*, and it plays a key role in the coupled cluster formalism.*

For clarity of exposition, we will adopt the convention of denoting by small bold letters such as $\boldsymbol{r}, \boldsymbol{s}, \boldsymbol{t}$*, and* $\boldsymbol{w}$*, etc., elements of the coefficient space* $\mathbb{V}$ *and denoting by capital curly letters such as* $\mathcal{R}, \mathcal{S}, \mathcal{T}$*, and* $\mathcal{W}$*, etc., the corresponding cluster operators with the understanding that* $\mathcal{R} := \sum_{\mu\in\mathcal{I}} \boldsymbol{r}_\mu \mathcal{X}_\mu$*,* $\mathcal{S} := \sum_{\mu\in\mathcal{I}} \boldsymbol{s}_\mu \mathcal{X}_\mu$*, and so on.*

Remark 2.3.5 (Representation of Elements of the Complementary Subspace $\widetilde{\mathcal{V}}$)**.**

Consider the setting of Definition 2.3.6 and recall Definition 2.3.4 of the space $\widetilde{\mathcal{V}} \subset \widehat{\mathcal{H}}^1$*. It is not difficult to see that every element* $\Phi_{\boldsymbol{s}} := \sum_{\mu\in\mathcal{I}} \boldsymbol{s}_\mu \mathcal{X}_\mu \Psi_0 \in \widetilde{\mathcal{V}}$ *generates a sequence* $\boldsymbol{s} := \{\boldsymbol{s}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$ *such that*

$$\Phi_{\boldsymbol{s}} = \sum_{\mu\in\mathcal{I}} \boldsymbol{s}_\mu \mathcal{X}_\mu \Psi_0 = \mathcal{S}\Psi_0.$$

Conversely, given any sequence $\mathbf{w} := \{\boldsymbol{w}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$*, we can define the function* $\Phi_{\boldsymbol{w}} \in \widetilde{\mathcal{V}}$ *as*

$$\Phi_{\boldsymbol{w}} = \sum_{\mu\in\mathcal{I}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0 = \mathcal{W}\Psi_0.$$

Therefore, in the sequel (in particular in Chapter 2.4), we will occasionally write elements of the space $\widetilde{\mathcal{V}}$ *as, for instance,* $\mathcal{S}\Psi_0$ *or* $\mathcal{W}\Psi_0$ *where* $\mathcal{S} := \sum_{\mu\in\mathcal{I}} \boldsymbol{s}_\mu \mathcal{X}_\mu$ *and* $\mathcal{W} := \sum_{\mu\in\mathcal{I}} \boldsymbol{w}_\mu \mathcal{X}_\mu$ *for some sequences* $\boldsymbol{s} := \{\boldsymbol{s}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$ *and* $\boldsymbol{w} := \{\boldsymbol{w}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$.

The following theorem now summarises the main properties of the excitation operators $\mathcal{X}_\mu$, $\mu \in \mathcal{I}$ and cluster operators constructed from these excitation operators. The establishment of these properties in infinite dimensions was the main achievement of the article [82]. In finite-dimensions, where the situation is considerably simpler from a topological point of view, these results were first proven in the mathematical literature in [87].

Theorem 2.3.3 (Properties of Excitation and Cluster Operators)**.**
Let the excitation index set $\mathcal{I}$ *be defined through Definition 2.3.1, let the excitation operators* $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{I}}$ *and de-excitation operators* $\{\mathcal{X}_\mu^\dagger\}_{\mu\in\mathcal{I}}$ *be defined through Definitions 2.3.2 and 2.3.3 respectively, and let the Hilbert space* $\mathbb{V}$ *of sequences be defined through Definition 2.3.6. Then*

1. *For all* $\mu, \nu \in \mathcal{I}$*, it holds that* $\mathcal{X}_\mu \mathcal{X}_\nu = \mathcal{X}_\nu \mathcal{X}_\mu$ *and* $\mathcal{X}_\mu^\dagger \mathcal{X}_\nu^\dagger = \mathcal{X}_\nu^\dagger \mathcal{X}_\mu^\dagger$.

2. *For every* $\Phi \in \widehat{\mathcal{H}}^1$ *that satisfies the so-called intermediate normalisation condition* $(\Phi, \Psi_0)_{\mathcal{L}^2} = 1$*, there exists a unique sequence* $\boldsymbol{r} = \{\boldsymbol{r}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$ *with corresponding cluster operator* $\mathcal{R} = \sum_{\mu\in\mathcal{I}} \boldsymbol{r}_\mu \mathcal{X}_\mu$ *such that*

$$\Phi = \Psi_0 + \mathcal{R}\Psi_0.$$

3. *Let* $\boldsymbol{t} \in \mathbb{V}$*. Then*

   - *The cluster operator* $\mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu$ *is a bounded linear map from* $\widehat{\mathcal{H}}^1$ *to* $\widehat{\mathcal{H}}^1$ *and there exists a constant* $\beta > 0$ *depending only on* Q *such that*

$$\|\boldsymbol{t}\|_{\mathbb{V}} \leqslant \|\mathcal{T}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1} \leqslant \beta \|\boldsymbol{t}\|_{\mathbb{V}}.$$

   - *The de-excitation cluster* $\mathcal{T}^\dagger = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu^\dagger$ *is also bounded linear map from* $\widehat{\mathcal{H}}^1$ *to* $\widehat{\mathcal{H}}^1$ *and there exists a constant* $\beta^\dagger > 0$ *depending only on* N *such that*

$$\|\mathcal{T}^\dagger\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1} \leqslant \beta^\dagger \|\boldsymbol{t}\|_{\mathbb{V}}.$$

   - *The cluster operator* $\mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu$ *has an extension to a bounded linear operator from* $\widehat{\mathcal{H}}^{-1}$ *to* $\widehat{\mathcal{H}}^{-1}$*.*

4. *Define the set of operators*

$$\mathfrak{L} := \left\{ t_0 \mathrm{I} + \mathcal{T} : \quad t_0 \in \mathbb{R},\ \mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu \quad \textit{such that} \quad \boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V} \right\}$$

   The following hold:

- *The set $\mathfrak{L}$ forms a closed commutative subalgebra in the algebra of bounded linear operators acting from $\widehat{\mathscr{H}}^1$ to $\widehat{\mathscr{H}}^1$ (and also from $\widehat{\mathscr{H}}^{-1}$ to $\widehat{\mathscr{H}}^{-1}$).*
- *The subalgebra $\mathfrak{L}$ is closed under inversion and the spectrum of any $\mathfrak{L} \ni \mathscr{L} = t_0\mathrm{I} + \mathcal{T}$ is exactly $\sigma(\mathscr{L}) = \{t_0\}$.*
- *Any element in $\mathfrak{L}$ of the form $\mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu$ with $\boldsymbol{t} \in \mathbb{V}$ is nilpotent: it holds that $\mathcal{T}^{Q+1} \equiv 0$.*
- *The exponential function is a locally $\mathscr{C}^\infty$ map on $\mathfrak{L}$, and is a bijection from the sub-algebra*

$$\left\{ \mathcal{T} : \quad \mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu \quad \textit{such that} \quad \boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V} \right\}.$$

to the sub-algebra

$$\left\{ \mathrm{I} + \mathcal{T} : \quad \mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu \quad \textit{such that} \quad \boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V} \right\}.$$

As a consequence of Theorem 2.3.3 (c.f., Properties (*2*) and (*4*)), one can prove that any intermediately normalised element of the Q-particle space can be parameterised through an exponential cluster operator. More precisely, given the excitation index set $\mathcal{I}$ defined through Definition 2.3.1 and the excitation operators $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{I}}$ defined through Definition 2.3.2, for any $\Phi \in \widehat{\mathscr{H}}^1$ such that $(\Phi, \Psi_0)_{\mathcal{L}^2} = 1$, there exists a unique sequence $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$ and a unique cluster operator $\mathcal{T} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}_\mu \mathcal{X}_\mu$ such that

$$\Phi = e^{\mathcal{T}} \Psi_0. \tag{2.12}$$

A proof of this statement in the infinite-dimensional setting can be found in [82] while the corresponding proof for the finite-dimensional case is given in [87].

Equation (2.12) implies in particular that if the sought-after ground state wave-function $\Psi^*_{\mathrm{GS}} \in \widehat{\mathscr{H}}^1$ that solves the minimisation problem (2.3a) is intermediately normalised, then it can also be written in the form

$$\Psi^*_{\mathrm{GS}} = e^{\mathcal{T}^*} \Psi_0,$$

for some sequence $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{I}}$ and corresponding cluster operator $\mathcal{T}^* = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu$. In other words, the minimisation problem (2.3a) can be replaced by an equivalent problem which consists of finding the sequence $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$ used to construct the appropriate cluster operator $\mathcal{T}^*$ that appears in the exponential parametrisation of Ψ^*. Indeed, it follows from the definition of Ψ^*_{GS} and such an exponential cluster operator $e^{\mathcal{T}^*}$ that

$$\mathcal{E}^*_{\mathrm{GS}} e^{\mathcal{T}^*} \Psi_0 = \mathcal{E}^*_{\mathrm{GS}} \Psi^*_{\mathrm{GS}} = H\Psi^*_{\mathrm{GS}} = H e^{\mathcal{T}^*} \Psi_0, \quad \text{and therefore} \quad \mathcal{E}^*_{\mathrm{GS}} \Psi_0 = e^{-\mathcal{T}^*} H e^{\mathcal{T}^*} \Psi_0.$$

Recalling now that for any excitation index $\mu \in \mathcal{I}$, the excited determinant $\Psi_\mu = \mathcal{X}_\mu \Psi_0$ is $\widehat{\mathcal{L}}^2$-orthogonal to the reference determinant Ψ_0, we are led to the *continuous* coupled cluster equations.

Continuous Coupled Cluster Equations:

Let the excitation index set $\mathcal{I}$ be defined through Definition 2.3.1 and let the excitation operators $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{I}}$ be defined through Definition 2.3.2. We seek a sequence $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$ such that for all $\nu \in \mathcal{I}$ we have

$$\left\langle \mathcal{X}_\nu \Psi_0, e^{-\mathcal{T}^*} H e^{\mathcal{T}^*} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = 0, \quad \text{where } \mathcal{T}^* = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu. \tag{2.13}$$

Once Equation (2.13) has been solved, the associated coupled cluster energy $\mathcal{E}^*_{\mathrm{CC}}$ is given by

$$\mathcal{E}^*_{\mathrm{CC}} := \left\langle \Psi_0, e^{-\mathcal{T}^*} H e^{\mathcal{T}^*} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}, \quad \text{where } \mathcal{T}^* = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu. \tag{2.14}$$

Remark 2.3.6 (Solutions to the Continuous Coupled Cluster Equations)**.**
Consider the continuous coupled cluster equations (2.13)*. Under the assumption that the ground state wave-function* Ψ_{GS} *of the electronic Hamiltonian* $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ *is intermediately normalisable with respect to the chosen reference determinant* Ψ_0*, i.e., it is not orthogonal to* Ψ_0*, it is obvious that there exists a corresponding solution to this non-linear system of equations. Indeed, by Equation* (2.12)*, there exists a sequence* $\boldsymbol{t}^*_{\mathrm{GS}} = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{I}} \in \mathbb{V}$ *such that* $\Psi^*_{\mathrm{GS}} = e^{\mathcal{T}^*_{\mathrm{GS}}}\Psi_0$ *with* $\mathcal{T}^*_{\mathrm{GS}} = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu$*, and it can readily be verified that this sequence* $\boldsymbol{t}^*_{\mathrm{GS}}$ *solves exactly Equation* (2.13)*, and consequently* $\mathcal{E}^*_{\mathrm{CC}} = \mathcal{E}^*_{\mathrm{GS}}$*.*

Of course, $\boldsymbol{t}^*_{\mathrm{GS}}$ *defined as above need not be the unique solution to the coupled cluster equations* (2.13)*. In fact, as we discuss in the next Chapter 2.4,* every *intermediately normalisable eigenfunction of the electronic Hamiltonian will generate a solution to Equation* (2.13)*. From a theoretical point of view, this means that only* local *well-posedness results can be expected to hold for the continuous CC equations.*

The continuous coupled cluster equations are an infinite system of non-linear equations and thus cannot be solved exactly. Instead, one introduces an approximation of the continuous coupled cluster equations by considering, instead of the global excitation index $\mathcal{I}$, some finite subset $\mathcal{I}_h \subset \mathcal{I}$ and solving only the equations associated with this subset of excitation indices. This procedure results in the so-called discrete coupled cluster equations.

Discrete Coupled Cluster Equations:
Let the excitation index set $\mathcal{I}$ be defined through Definition 2.3.1, let $\mathcal{I}_h \subset \mathcal{I}$ denote any *finite* subset of excitation indices, and let the excitation operators $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{I}}$ be defined through Definition 2.3.2. We seek a coefficient vector $\boldsymbol{t}^*_h = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{I}_h} \in \ell^2(\mathcal{I}_h)$ such that for all $\nu \in \mathcal{I}_h$ we have

$$\left\langle \mathcal{X}_\nu \Psi_0, e^{-\mathcal{T}^*_h} H e^{\mathcal{T}^*_h} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = 0, \quad \text{where } \mathcal{T}^*_h = \sum_{\mu\in\mathcal{I}_h} \boldsymbol{t}^*_\mu \mathcal{X}_\mu. \tag{2.15}$$

The associated discrete ground state energy $\mathcal{E}^*_{h,\mathrm{CC}}$ is given by

$$\mathcal{E}^*_{h,\mathrm{CC}} := \left\langle \Psi_0, e^{-\mathcal{T}^*_h} H e^{\mathcal{T}^*_h} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}, \quad \text{where } \mathcal{T}^*_h = \sum_{\mu\in\mathcal{I}_h} \boldsymbol{t}^*_\mu \mathcal{X}_\mu. \tag{2.16}$$

The discrete CC equations will be the subject of further discussion in Chapter 2.5 where we will analyse their well-posedness for some specific choices of the excitation index subsets and Q-particle basis sets. For the moment, we conclude this section with a remark on the nature of the solutions to these discrete equations.

Remark 2.3.7 (Solutions to the Discrete Coupled Cluster Equations)**.**

Consider the discrete coupled cluster equations (2.15)*. As in the continuous case, there is a priori no reason for solutions of Equation* (2.15) *to be globally unique. Indeed, numerical experience confirms that solutions to Equation* (2.15) *are very often* not *unique (see, e.g., [111, 78, 75, 59, 77]). Nevertheless, in practice the discrete CC equations are solved very frequently by the quantum chemical community when performing electronic structure calculations, usually using some type of iterative Newton method, and it is* hoped *that if one starts from a sufficiently accurate initial point, then the resulting solution* $\boldsymbol{t}_h^* \in \ell^2(\mathcal{G}_h)$ *of Equation* (2.15) *approximates, in some sense, an exact solution* $\boldsymbol{t}^*$ *of the continuous CC equations* (2.13) *that generates the intermediately normalised ground state wave-function. Of course there are no mathematical guarantees that this procedure works, and the current reputation of coupled cluster methods as a 'gold-standard' in computational quantum chemistry seems to be based mostly on successful empirical experience.*

Having introduced the continuous and discrete coupled cluster equations, the remainder of this part will be concerned with their (local) well-posedness analysis. We will first analyse the continuous coupled cluster equations (2.13) in Chapter 2.4, following which we will study a particular class of the discrete coupled cluster equations (2.15) in Chapter 2.5.

2.4 Well-posedness of the continuous coupled cluster equations

Throughout this section, we assume the setting of Chapters 2.2 and 2.3, and we recall in particular the notion of excitation operators and the continuous coupled cluster equations. We begin by defining the so-called coupled cluster function, which will be the main object of study in this section.

Definition 2.4.1 (Coupled Cluster function)**.**

Let the excitation index set $\mathcal{G}$ *be defined through Definition 2.3.1 and let the excitation operators* $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{G}}$ *be defined through Definition 2.3.2. We define the coupled cluster function* $f\colon \mathbb{V} \to \mathbb{V}^*$ *as the mapping with the property that for all* $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{G}}, \boldsymbol{s} = \{\boldsymbol{s}_\nu\}_{\nu\in\mathcal{G}} \in \mathbb{V}$ *it holds that*

$$\big\langle \boldsymbol{s}, f(\boldsymbol{t})\big\rangle_{\mathbb{V}\times\mathbb{V}^*} := \left\langle \sum_{\nu\in\mathcal{G}} \boldsymbol{s}_\nu \mathcal{X}_\nu \Psi_0, e^{-\mathcal{T}} H e^{\mathcal{T}} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}, \qquad \text{where } \mathcal{T} = \sum_{\mu\in\mathcal{G}} \boldsymbol{t}_\mu \mathcal{X}_\mu.$$

Remark 2.4.1 (Justification of the Domain and Range of Coupled Cluster Function)**.**

Consider Definition 2.4.1 of the coupled cluster function. The fact that f *is indeed a mapping from* $\mathbb{V}$ *to* $\mathbb{V}^*$ *is a direct consequence of the boundedness of the electronic Hamiltonian* $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ *and the exponential cluster operators* $e^{\mathcal{T}} : \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ *and* $e^{-\mathcal{T}} : \widehat{\mathcal{H}}^{-1} \to \widehat{\mathcal{H}}^{-1}$*. Indeed, for all*

$\boldsymbol{s} = \{\boldsymbol{s}_\nu\}_{\nu\in\mathcal{J}} \in \mathbb{V}$ *and all* $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{J}} \in \mathbb{V}$ *it holds that*

$$\begin{aligned}\left|\left\langle \boldsymbol{s}, f(\boldsymbol{t})\right\rangle_{\mathbb{V}\times\mathbb{V}^*}\right| &= \left|\left\langle \sum_{\nu\in\mathcal{J}} \boldsymbol{s}_\nu \mathcal{X}_\nu \Psi_0, e^{-\mathcal{T}} H e^{\mathcal{T}} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right| \\ &\leqslant \left\| \sum_{\nu\in\mathcal{J}} \boldsymbol{s}_\nu \mathcal{X}_\nu \Psi_0 \right\|_{\widehat{\mathcal{H}}^1} \left\| e^{-\mathcal{T}} H e^{\mathcal{T}} \Psi_0 \right\|_{\widehat{\mathcal{H}}^{-1}} \\ &\leqslant \|\boldsymbol{s}\|_{\mathbb{V}} \|e^{-\mathcal{T}}\|_{\widehat{\mathcal{H}}^{-1}\to\widehat{\mathcal{H}}^{-1}} \|H\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^{-1}} \|e^{\mathcal{T}}\Psi_0\|_{\widehat{\mathcal{H}}^1}.\end{aligned}$$

Equipped with Definition 2.4.1 of the coupled cluster function, let us point out that the continuous coupled cluster equations (2.13) can be re-written in the following weak form.

Weak Form of the Continuous Coupled Cluster equations:
Let the excitation index set $\mathcal{J}$ be defined through Definition 2.3.1, let the Hilbert space of sequences $\mathbb{V} \subset \ell^2(\mathcal{J})$ be defined through Definition 2.3.6, and let the coupled cluster function $f\colon \mathbb{V} \to \mathbb{V}^*$ be defined through Definition 2.4.1. We seek a sequence $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{J}} \in \mathbb{V}$ such that for all sequences $\boldsymbol{s} = \{\boldsymbol{s}_\nu\}_{\nu\in\mathcal{J}} \in \mathbb{V}$ it holds that

$$\langle \boldsymbol{s}, f(\boldsymbol{t})\rangle_{\mathbb{V}\times\mathbb{V}^*} = 0. \tag{2.17}$$

As we shall see in Chapter 2.5, this point of view will allow us to interpret the *truncated* coupled cluster equations (2.15) as Galerkin discretisations of Equation (2.17), which will be useful for the purpose of the numerical analysis.

The following extremely significant result, proven in [82], establishes a precise relationship between zeros of the coupled cluster function defined through Definition 2.4.1 (i.e., solutions of the continuous CC equations (2.17)) and intermediately normalised eigenfunctions of the electronic Hamiltonian defined through Equation (2.2).

Theorem 2.4.1 (Relation between Coupled Cluster Zeros and Eigenfunctions of Electronic Hamiltonian)**.**

Let the coupled cluster function $f\colon \mathbb{V} \to \mathbb{V}^*$ *be defined through Definition 2.4.1 and let the electronic Hamiltonian be given by Equation* (2.2)*. Then*

1. *For any zero* $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{J}} \in \mathbb{V}$ *of the CC function, the function* $\Psi^* = e^{\mathcal{T}^*}\Psi_0 \in \widehat{\mathcal{H}}^1$ *with* $\mathcal{T}^* = \sum_{\mu\in\mathcal{J}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu$ *is an intermediately normalised eigenfunction of the electronic Hamiltonian. Moreover, the eigenvalue corresponding to the eigenfunction* Ψ^* *coincides with the CC energy* $\mathcal{E}^*_{\mathrm{CC}}$ *generated by* $\boldsymbol{t}^*$ *as defined through Equation* (2.14)*.*

2. *Conversely, for any intermediately normalised eigenfunction* $\Psi^* \in \widehat{\mathcal{H}}^1$ *of the electronic Hamiltonian, there exists* $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{J}} \in \mathbb{V}$ *such that* $\boldsymbol{t}^*$ *is a zero of the CC function and* $\Psi^* = e^{\mathcal{T}^*}\Psi_0 \in \widehat{\mathcal{H}}^1$ *with* $\mathcal{T}^* = \sum_{\mu\in\mathcal{J}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu$*. Moreover, the CC energy* $\mathcal{E}^*_{\mathrm{CC}}$ *generated by* $\boldsymbol{t}^*$ *as defined through Equation* (2.14) *coincides with the eigenvalue corresponding to the eigenfunction* Ψ^**.*

In other words every intermediately normalisable eigenfunction of the electronic Hamiltonian (2.2) corresponds to a zero of the coupled cluster function defined through Definition 2.4.1 and vice-versa. The goal of our analysis in this section is to study the nature of these zeros of the coupled cluster function and, in particular, to derive sufficient conditions that guarantee the *simplicity* of the zeros. Indeed, if we know that some $\boldsymbol{t}^* \in \mathbb{V}$ is a simple zero of the coupled

cluster function, then this will allow us to deduce *local invertibility* of the coupled cluster function at $\boldsymbol{t}^* \in \mathbb{V}$ and thereby derive both *local uniqueness* and *local residual-based* error estimates for the CC equations (2.17). Arguments of this nature are standard in the literature on non-linear numerical analysis (see, e.g., [96, Proposition 2.1] or [12, Theorem 2.1]) and are usually based on the invertibility of the Fréchet derivative of the non-linear function being studied. The next step in our analysis therefore will be to study carefully the Fréchet derivative of the coupled cluster function. Before proceeding with this analysis however, let us comment on the existing numerical analysis of the coupled cluster equation (2.17).

Remark 2.4.2 (Existing Approaches in the Numerical Analysis of the CC equations (2.17))**.**

The existing literature on the numerical analysis of coupled cluster methods is rather sparse. The first numerical analysis of the single reference coupled cluster– in the finite-dimensional setting– is due to R. Schneider in [87]. The analysis carried out in [87] was then extended to the infinite-dimensional setting (as considered here) in the subsequent articles [82] and [83]. The former article showed that the mathematical objects used to formulate the coupled cluster method (such as excitation operators) are bounded operators on appropriate infinite-dimensional Hilbert spaces so that the coupled cluster equations can be stated in infinite-dimensions (prior to this article, the CC equations were always *written in a finite-dimensional setting). The article [83] used these tools and the ideas developed in [87] to perform a numerical analysis of the infinite-dimensional coupled cluster equations. Additional articles on the mathematical analysis of CC methods have since appeared, including [61] which studies the so-called extended coupled cluster method, [42] which studies the so-called tailored coupled cluster method, [30] which studies the finite-dimensional CC equations using topological degree theory, and [43] which analyses the root structure of the CC equations using tools from algebraic geometry.*

The aforementioned articles have two important features in common: First they are concerned with the (local) analysis of the 'ground-state' zero of the coupled cluster function, i.e., with the zero $\boldsymbol{t}^*_{\mathrm{GS}}$ such that $e^{\mathcal{T}^*_{\mathrm{GS}}}\Psi_0 = \Psi^*_{\mathrm{GS}}$. This of course makes sense since the vast majority of coupled cluster calculations are targeted at approximating the ground state energy of the electronic Hamiltonian.

Second and more importantly, the well-posedness analysis in all of the above articles is based on proving a local, strong monotonicity *property of the coupled cluster function at $\boldsymbol{t}^*_{\mathrm{GS}}$. Taking the example of the article [83] whose notation closely aligns with ours, let the coupled cluster function $f\colon \mathbb{V} \to \mathbb{V}^*$ be defined through Definition 2.4.1. Then it is shown in [83] that for $\delta > 0$ sufficiently small, there exists a constant Γ such that for all $\boldsymbol{w}, \boldsymbol{s} \in \mathrm{B}_\delta(\boldsymbol{t}^*_{\mathrm{GS}}) \subset \mathbb{V}$ it holds that*

$$\langle \boldsymbol{w} - \boldsymbol{s}, f(\boldsymbol{w}) - f(\boldsymbol{s})\rangle_{\mathbb{V}\times\mathbb{V}^*} \geqslant \Gamma \|\boldsymbol{w} - \boldsymbol{s}\|_{\mathbb{V}}. \tag{2.18}$$

If the constant Γ can be shown to be strictly positive, then the local monotonicity property (2.18) *immediately yields local well-posedness of both the continuous coupled cluster equations as well as sufficiently rich Galerkin discretisations thereof*[3]*. Quasi-optimal error estimates for the CC energy can then also be derived using the dual weighted residual approach developed by Rannacher et al. [4, Chapter 6] .*

The main drawback of the above approach is that the actual local monotonicity constant Γ derived from this analysis (see [83, Theorem 3.4]) is of the form:

$$\begin{aligned}\Gamma \quad &= \omega\gamma - \left\|\mathcal{J}^*_{\mathrm{GS}} - (\mathcal{J}^*_{\mathrm{GS}})^\dagger\right\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1} \ \|H - \mathcal{E}^*_{\mathrm{GS}}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^{-1}} - \mathcal{O}\left(\|\boldsymbol{t}^*_{\mathrm{GS}}\|^2_{\mathbb{V}}\right) &\text{(2.19a)}\\ &\geqslant \omega\gamma - \left(\beta + \beta^\dagger\right)\|\boldsymbol{t}^*_{\mathrm{GS}}\|_{\mathbb{V}}\,\|H - \mathcal{E}^*_{\mathrm{GS}}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^{-1}} - \mathcal{O}\left(\|\boldsymbol{t}^*_{\mathrm{GS}}\|^2_{\mathbb{V}}\right) &\text{(2.19b)}\end{aligned}$$

[3] Although this point of view is not taken in these articles, the local monotonicity condition (2.18) essentially corresponds to proving that the coupled cluster Fréchet derivative at $\boldsymbol{t}^*_{\mathrm{GS}}$ is a positive definite operator.

where $\gamma > 0$ denotes the coercivity constant of the shifted electronic Hamiltonian $H - \mathcal{E}^*_{\mathrm{GS}}$ on $\{\Psi^*_{\mathrm{GS}}\}^\perp$, the constant $\omega \in (0,1)$ is a prefactor depending on $\|\Psi_0 - \Psi^*_{\mathrm{GS}}\|_{\widehat{\mathcal{L}}^2}$, and $\beta, \beta^\dagger$ are the continuity constants of the mappings $\mathbb{V} \ni \boldsymbol{t} \mapsto \mathcal{T} \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ and and $\mathbb{V} \ni \boldsymbol{t} \mapsto \mathcal{T}^\dagger \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ respectively as given in Theorem 2.3.3.

Consequently, the constant Γ is positive provided that $\|\boldsymbol{t}^*_{\mathrm{GS}}\|_{\mathbb{V}}$ is small enough. However, according to the theoretical analysis in [82], the constants $\beta, \beta^\dagger$ grow combinatorially *in the number of electrons Q in the system, and thus as soon as $Q \approx 10$ or larger, the lower bound* (2.19b) *for the constant Γ is no longer positive. Similar issues arise in the local monotonicity constants derived in the other articles [42] and [61].*

To make matters worse, even if we rely on the sharper Inequality (2.19a)*, numerical experiments involving small, relatively well-behaved molecules for which it is well-known (from numerical experience) that the coupled cluster method works well, reveal that (see Table 2.1 below)*

$$\left\|\mathcal{T}^*_{\mathrm{GS}} - (\mathcal{T}^*_{\mathrm{GS}})^\dagger\right\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1} \; \|H - \mathcal{E}^*_{\mathrm{GS}}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}} > \gamma.$$

In other words, the assumptions required to establish local strong monotonicity of the coupled cluster function, namely, smallness of the amplitude vector norm $\|\boldsymbol{t}^*_{\mathrm{GS}}\|_{\mathbb{V}}$ seem restrictive and not satisfied in many practical examples. As a consequence, the hope of obtaining quantitative *a posteriori error estimates for the coupled cluster equations appears difficult to achieve.*

Molecule	Coercivity constant γ	$\|\boldsymbol{t}^*_{\mathrm{FCI}}\|_{\mathbb{V}}$	Monotonicity constant Γ from Eq. (2.19a)	Hartree-Fock Energy Error (Hartree)	CCSD Energy Error (Hartree)
BeH2	0.3257	0.2343	0.0363	3.50×10^{-2}	3.83×10^{-4}
BH3	0.2903	0.2844	-0.0950	5.40×10^{-2}	3.74×10^{-4}
HF	0.3010	0.2038	-0.0083	2.81×10^{-2}	3.02×10^{-5}
H2O	0.3471	0.2687	0.0249	5.01×10^{-2}	1.18×10^{-4}
LiH	0.2617	0.1792	-0.0065	2.04×10^{-2}	1.14×10^{-5}
NH3	0.3868	0.3074	-0.0325	6.61×10^{-2}	2.18×10^{-4}

Table 2.1: Examples of numerically computed local monotonicity constants for a collection of small molecules at equilibrium geometries. The calculations were performed in STO-6G basis sets with the exception of the HF and LiH molecules for which 6-31G basis sets were used. In all cases, the Full-CI solution was taken as the reference solution. To simplify calculations, the canonical $\widehat{\mathcal{H}}^1$ norm was replaced with an equivalent norm induced by the mean-field Hartree Fock operator (see, e.g., [87]).

We begin our analysis with the following proposition whose essence seems known (c.f., [87, Theorem 4.16], [83, Lemma 3.1] and [29, Lemma 4.6]) but that has not been expressed in the current form in the existing literature.

Proposition 2.4.1 (Coupled Cluster Fréchet Derivative)**.**

Let the excitation index set $\mathcal{I}$ be defined through Definition 2.3.1, let the excitation operators $\{\mathcal{X}_\mu\}_{\mu \in \mathcal{I}}$ be defined through Definition 2.3.2, and let the coupled cluster function $f \colon \mathbb{V} \to \mathbb{V}^*$ be defined through Definition 2.4.1. Then,

- *For any* $\boldsymbol{t} = \{\boldsymbol{t}_\mu\}_{\mu\in\mathcal{G}} \in \mathbb{V}$*, the Fréchet derivative* $\mathrm{D}f(\boldsymbol{t})\colon \mathbb{V} \to \mathbb{V}^*$ *of the coupled cluster function* f *at* $\boldsymbol{t}$ *is the mapping with the property that for all* $\boldsymbol{s}, \boldsymbol{w} \in \mathbb{V}$ *with* $\boldsymbol{s} = \{\boldsymbol{s}_\nu\}_{\nu\in\mathcal{G}}$ *and* $\boldsymbol{w} = \{\boldsymbol{w}_\mu\}_{\mu\in\mathcal{G}}$ *it holds that*

$$\langle \boldsymbol{w}, \mathrm{D}f(\boldsymbol{t})\boldsymbol{s}\rangle_{\mathbb{V}\times\mathbb{V}^*} = \left\langle \sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}} \left[H, \sum_{\nu\in\mathcal{G}} \boldsymbol{s}_\nu \mathcal{X}_\nu\right] e^{\mathcal{T}} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}, \tag{2.20}$$

where $[\cdot,\cdot]$ *denotes the commutator and* $\mathcal{T} := \sum_{\mu\in\mathcal{G}} \boldsymbol{t}_\mu \mathcal{X}_\mu$.

- $f\colon \mathbb{V} \to \mathbb{V}^*$ *is a* $\mathscr{C}^\infty$ *mapping.*

Proof. We start with the proof of the first assertion. This portion of the proof will proceed in two steps:

- We will obtain an expression for the Gateaux derivative $\mathrm{D}f(\boldsymbol{t}),\ t \in \mathbb{V}$ of the coupled cluster function f, and we will show that this agrees with the expression offered by Equation (2.20).
- We will show that the Gateaux derivative is continuous as a function of $\boldsymbol{t}$, i.e., the mapping $\boldsymbol{t} \mapsto \mathrm{D}f(t)\colon \mathbb{V} \to \mathbb{V}^*$ is continuous.

Let $\boldsymbol{t}, \boldsymbol{s} \in \mathbb{V}$ be arbitrary. Thanks to Remark 2.4.1, we observe that for any $h \geqslant 0$ it holds that $f(\boldsymbol{t} + h\boldsymbol{s}) \in \mathbb{V}^*$. It follows that for any $h > 0$ and any $\boldsymbol{w} \in \mathbb{V}$ we have that

$$\begin{aligned}\langle \boldsymbol{w}, f(\boldsymbol{t}+h\boldsymbol{s}) - f(\boldsymbol{t})\rangle_{\mathbb{V}\times\mathbb{V}^*} &= \left\langle \sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, \left(e^{-\mathcal{T}-h\mathcal{S}} H e^{\mathcal{T}+h\mathcal{S}} - e^{-\mathcal{T}} H e^{\mathcal{T}}\right)\Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \\ &= \left\langle \sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}} \left(e^{-h\mathcal{S}} H e^{h\mathcal{S}} - H\right) e^{\mathcal{T}} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}},\end{aligned}$$

where we have denoted $\mathcal{S} := \sum_{\mu\in\mathcal{G}} \boldsymbol{s}_\mu \mathcal{X}_\mu$ and we have used the fact that $\mathcal{T}$ and $\mathcal{S}$ commute (see the first assertion of Theorem 2.3.3).

As a consequence, using once again the commutativity of $\mathcal{T}$ and $\mathcal{S}$ together with the power series expansion of the exponential cluster operator, we deduce that

$$\begin{aligned}\lim_{h\to 0} \frac{\langle \boldsymbol{w}, f(\boldsymbol{t}+h\boldsymbol{s}) - f(\boldsymbol{t})\rangle_{\mathbb{V}\times\mathbb{V}^*}}{h} &= \left\langle \sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}} \left(-\mathcal{S}H + H\mathcal{S}\right) e^{\mathcal{T}} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \\ &= \left\langle \sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}} \left[H, \sum_{\nu\in\mathcal{G}} \boldsymbol{s}_\nu \mathcal{X}_\nu\right] e^{\mathcal{T}} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}.\end{aligned} \tag{2.21}$$

In order to show that the expression offered by Equation (2.21) defines the Gateaux derivative $\mathrm{D}f(t)\colon \mathbb{V} \to \mathbb{V}^*$, we must show that this operator is bounded. Recalling from Definition 2.3.4, the subspace $\widetilde{\mathcal{V}} \subset \widehat{\mathcal{H}}^1$ as the orthogonal complement of $\{\Psi_0\}$, let us therefore define $\mathcal{A}(\boldsymbol{t})\colon \widetilde{\mathcal{V}} \to \widehat{\mathcal{H}}^{-1}$ as

$$\forall \Phi \in \widehat{\mathcal{H}}^1,\ \ \forall \mathcal{S}\Psi_0 \in \widetilde{\mathcal{V}} \ \text{ with } \mathcal{S} = \sum_\mu \boldsymbol{s}_\mu \mathcal{X}_\mu\colon \quad \langle \Phi, \mathcal{A}(\boldsymbol{t})\mathcal{S}\Psi_0\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} := \left\langle \Phi, e^{-\mathcal{T}}[H,\mathcal{S}]e^{\mathcal{T}}\Psi_0\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}.$$

We claim that $\mathcal{A}(\boldsymbol{t})$ defines a bounded linear operator. Indeed, a direct calculation shows

that $\forall \Phi \in \widehat{\mathscr{H}}^1$ and $\forall \mathcal{S}\Psi_0 \in \widetilde{\mathcal{V}}$ with $\mathcal{S} = \sum_\mu \boldsymbol{s}_\mu \mathcal{X}_\mu$ we have

$$\begin{aligned}
\big|\, \langle \Phi, \mathcal{A}(\boldsymbol{t})\mathcal{S}\Psi_0\rangle_{\widehat{\mathscr{H}}^1\times\widehat{\mathscr{H}}^{-1}} \,\big| &= \Big|\big\langle \Phi, e^{-\mathcal{T}} H \mathcal{S} e^{\mathcal{T}}\Psi_0\big\rangle_{\widehat{\mathscr{H}}^1\times\widehat{\mathscr{H}}^{-1}} - \big\langle \Phi, e^{-\mathcal{T}}\mathcal{S}He^{\mathcal{T}}\Psi_0\big\rangle_{\widehat{\mathscr{H}}^1\times\widehat{\mathscr{H}}^{-1}}\Big| \\
&\leqslant \|\Phi\|_{\widehat{\mathscr{H}}^1}\, \big\|e^{-\mathcal{T}}\big\|_{\widehat{\mathscr{H}}^{-1}\to\widehat{\mathscr{H}}^{-1}}\, \|H\|_{\widehat{\mathscr{H}}^1\to\widehat{\mathscr{H}}^{-1}}\, \big\|e^{\mathcal{T}}\big\|_{\widehat{\mathscr{H}}^1\to\widehat{\mathscr{H}}^1}\, \|\mathcal{S}\Psi_0\|_{\widehat{\mathscr{H}}^1} \\
&\quad + \big\|\mathcal{S}^\dagger\Phi\big\|_{\widehat{\mathscr{H}}^1}\, \big\|e^{-\mathcal{T}}\big\|_{\widehat{\mathscr{H}}^{-1}\to\widehat{\mathscr{H}}^{-1}}\, \|H\|_{\widehat{\mathscr{H}}^1\to\widehat{\mathscr{H}}^{-1}}\, \big\|e^{\mathcal{T}}\Psi_0\big\|_{\widehat{\mathscr{H}}^1}\,,
\end{aligned}$$

where we have used the fact that the cluster operators $e^{\mathcal{T}}$ and $\mathcal{S}$ commute. Next, let us observe that by definition of the cluster operator $\mathcal{S}$ and the norm $\|\cdot\|_{\mathbb{V}}$, it holds that $\|\mathcal{S}\Psi_0\|_{\widehat{\mathscr{H}}^1} = \|\boldsymbol{s}\|_{\mathbb{V}}$. Consequently, recalling the continuity properties of cluster operators from Theorem 2.3.3 we deduce that

$$\big\|\mathcal{S}^\dagger\Phi\big\|_{\widehat{\mathscr{H}}^1} \leqslant \beta^\dagger \|\boldsymbol{s}\|_{\mathbb{V}} \|\Phi\|_{\widehat{\mathscr{H}}^1} = \beta^\dagger \|\mathcal{S}\Psi_0\|_{\widehat{\mathscr{H}}^1} \|\mathcal{W}\Psi_0\|_{\widehat{\mathscr{H}}^1},$$

where the constant $\beta^\dagger > 0$ is independent of $\mathcal{S}$.

Collecting terms now shows that $\mathcal{A}(\boldsymbol{t})\colon \widetilde{\mathcal{V}} \to \widehat{\mathscr{H}}^{-1}$ is indeed bounded, and therefore the Gateaux derivative $\mathrm{D}f(\boldsymbol{t})\colon \mathbb{V} \to \mathbb{V}^*$ is well-defined according to the expression offered by Equation (2.21). Since $\boldsymbol{t} \in \mathbb{V}$ was arbitrary, the coupled cluster function f is everywhere Gateaux differentiable.

It remains to prove that the Gateaux derivative $\mathrm{D}f(\boldsymbol{t})$ is in fact a Fréchet derivative. To this end, it suffices to show that the mapping $\mathbb{V} \ni \boldsymbol{t} \mapsto \mathrm{D}f(\boldsymbol{t})\colon \mathbb{V} \to \mathbb{V}^*$ is continuous. To do so, let $\{\boldsymbol{t}_n\}_{n\in\mathbb{N}} \subset \mathbb{V}$ be a sequence that converges to $\boldsymbol{t}$. It follows that

$$\lim_{n\to\infty} \|\mathrm{D}f(\boldsymbol{t}) - \mathrm{D}f(\boldsymbol{t}_n)\|_{\mathbb{V}\to\mathbb{V}^*} = \lim_{n\to\infty} \sup_{\substack{\boldsymbol{s}\in\mathbb{V}\\ \|\boldsymbol{s}\|_{\mathbb{V}}=1}} \sup_{\substack{\boldsymbol{w}\in\mathbb{V}\\ \|\boldsymbol{w}\|_{\mathbb{V}}=1}} |\, \langle \boldsymbol{w}, \mathrm{D}f(\boldsymbol{t})\boldsymbol{s} - \mathrm{D}f(\boldsymbol{t}_n)\boldsymbol{s}\rangle_{\mathbb{V}\times\mathbb{V}^*}\,|$$

$$= \lim_{n\to\infty} \sup_{\substack{\boldsymbol{s}\in\mathbb{V}\\ \|\boldsymbol{s}\|_{\mathbb{V}}=1}} \sup_{\substack{\boldsymbol{w}\in\mathbb{V}\\ \|\boldsymbol{w}\|_{\mathbb{V}}=1}} \left| \left\langle \sum_{\mu\in\mathcal{J}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}}\Big[H, \sum_{\nu\in\mathcal{J}} \boldsymbol{s}_\nu \mathcal{X}_\nu\Big] e^{\mathcal{T}}\Psi_0 - e^{-\mathcal{T}_n}\Big[H, \sum_{\nu\in\mathcal{J}} \boldsymbol{s}_\nu \mathcal{X}_\nu\Big] e^{\mathcal{T}_n}\Psi_0 \right\rangle_{\widehat{\mathscr{H}}^1\times\widehat{\mathscr{H}}^{-1}} \right|,$$

where for all $n \in \mathbb{N}$ we denote $\mathcal{T}_n := \sum_{\mu\in\mathcal{J}} (\boldsymbol{t}_n)_\mu \mathcal{X}_\mu$. Adding and subtracting suitable terms

yields the inequality

$$
\begin{aligned}
&\lim_{n\to\infty} \|\mathrm{D}f(t) - \mathrm{D}f(t_n)\|_{\mathbb{V}\to\mathbb{V}^*} \\
&\leqslant \lim_{n\to\infty} \sup_{\substack{\boldsymbol{s}\in\mathbb{V}\\ \|\boldsymbol{s}\|_{\mathbb{V}}=1}} \sup_{\substack{\boldsymbol{w}\in\mathbb{V}\\ \|\boldsymbol{w}\|_{\mathbb{V}}=1}} \left| \left\langle \sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, \left(e^{-\mathcal{T}} - e^{-\mathcal{T}_n}\right) \left[H, \sum_{\nu\in\mathcal{G}} \boldsymbol{s}_\nu \mathcal{X}_\nu\right] e^{\mathcal{T}} \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \right| \\
&+ \lim_{n\to\infty} \sup_{\substack{\boldsymbol{s}\in\mathbb{V}\\ \|\boldsymbol{s}\|_{\mathbb{V}}=1}} \sup_{\substack{\boldsymbol{w}\in\mathbb{V}\\ \|\boldsymbol{w}\|_{\mathbb{V}}=1}} \left| \left\langle \sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}_n} \left[H, \sum_{\nu\in\mathcal{G}} \boldsymbol{s}_\nu \mathcal{X}_\nu\right] \left(e^{\mathcal{T}} - e^{\mathcal{T}_n}\right) \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \right| \\
&\leqslant \lim_{n\to\infty} \sup_{\substack{\boldsymbol{s}\in\mathbb{V}\\ \|\boldsymbol{s}\|_{\mathbb{V}}=1}} \left\| \left(e^{-\mathcal{T}} - e^{-\mathcal{T}_n}\right) \left[H, \sum_{\nu\in\mathcal{G}} \boldsymbol{s}_\nu \mathcal{X}_\nu\right] e^{\mathcal{T}} \Psi_0 \right\|_{\widehat{\mathcal{H}}^{-1}} \\
&+ \lim_{n\to\infty} \sup_{\substack{\boldsymbol{s}\in\mathbb{V}\\ \|\boldsymbol{s}\|_{\mathbb{V}}=1}} \left\| e^{-\mathcal{T}_n} \left[H, \sum_{\nu\in\mathcal{G}} \boldsymbol{s}_\nu \mathcal{X}_\nu\right] \left(e^{\mathcal{T}} - e^{\mathcal{T}_n}\right) \Psi_0 \right\|_{\widehat{\mathcal{H}}^{-1}},
\end{aligned}
$$

where we have used the fact that $\|\boldsymbol{w}\|_{\mathbb{V}} = \left\|\sum_{\mu\in\mathcal{G}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0\right\|_{\widehat{\mathcal{H}}^1}$ by definition.

We can now use the fact that the exponential cluster operator is a locally $\mathscr{C}^\infty$ mapping on the algebra of cluster operators (see Theorem 2.3.3) together with the boundedness properties of the Hamiltonian H and excitation operators to deduce that both of the above limits are zero. Thus, $\lim_{n\to\infty} \|\mathrm{D}f(\boldsymbol{t}) - \mathrm{D}f(\boldsymbol{t}_n)\|_{\mathbb{V}\to\mathbb{V}^*} = 0$, which shows that $\mathrm{D}f(t)\colon \mathbb{V} \to \mathbb{V}^*$ as defined through Equation (2.20) is indeed the Fréchet derivative of the coupled cluster function $f\colon \mathbb{V} \to \mathbb{V}^*$ at $\boldsymbol{t} \in \mathbb{V}$.

In order to complete the proof of this proposition, we must demonstrate that the second assertion also holds, namely, that f is a $\mathscr{C}^\infty$ mapping from $\mathbb{V}$ to $\mathbb{V}^*$. To this end, it is sufficient to observe that higher order Gateaux derivatives of the coupled cluster function can be computed exactly as the first order Gateaux derivative given by Equation (2.21) with the single commutator being replaced by nested commutators. The fact that these Gateaux derivatives are also Fréchet derivatives is deduced in an identical fashion by making use of the fact that exponential cluster operator is a locally $\mathscr{C}^\infty$ map. This completes the proof. □

Proposition 2.4.1 has a number of important consequences that we now state. For the first result, let us recall from Theorem (2.4.1) that every zero $\boldsymbol{t}^* \in \mathbb{V}$ of the coupled cluster function is associated with an intermediately normalised eigenfunction $\Psi^* \in \widehat{\mathcal{H}}^1$ of the electronic Hamiltonian $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ defined through Equation (2.2).

Corollary 2.4.1.1 (Coupled Cluster Fréchet Derivative at Zeros of the Coupled Cluster Function)**.**

Let the excitation index set $\mathcal{I}$ be defined through Definition 2.3.1, let the excitation operators $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{G}}$ be defined through Definition 2.3.2, let the coupled cluster function $f\colon \mathbb{V} \to \mathbb{V}$ be defined through Definition 2.4.1, for any $\boldsymbol{t} \in \mathbb{V}$ let $\mathrm{D}f(\boldsymbol{t})$ denote the Fréchet derivative of the coupled cluster function as defined through Equation (2.20), let $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{G}} \in \mathbb{V}$ denote a zero of the CC function that generates the intermediately normalised eigenfunction $\Psi^* \in \widehat{\mathcal{H}}^1$ of the electronic Hamiltonian with corresponding eigenvalue $\mathcal{E}^*$. Then for all $\boldsymbol{s}, \boldsymbol{w} \in \mathbb{V}$ with $\boldsymbol{s} = \{\boldsymbol{s}_\nu\}_{\nu\in\mathcal{G}}$ and

$\boldsymbol{w} = \{\boldsymbol{w}_\mu\}_{\mu\in\mathcal{I}}$, *it holds that*

$$\langle \boldsymbol{w}, \mathrm{D}f(\boldsymbol{t}^*)\boldsymbol{s}\rangle_{\mathbb{V}\times\mathbb{V}^*} = \left\langle \sum_{\mu\in\mathcal{I}} \boldsymbol{w}_\mu \mathcal{X}_\mu \Psi_0, e^{-\mathcal{T}^*}\,(H-\mathcal{E}^*)\,e^{\mathcal{T}^*} \sum_{\nu\in\mathcal{I}} \boldsymbol{s}_\nu \mathcal{X}_\nu \Psi_0 \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \quad \text{where } \mathcal{T}^* := \sum_{\mu\in\mathcal{I}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu. \tag{2.22}$$

Proof. The proof follows by a direct calculation from Equation (2.20) by expanding the commutator, making use of the fact that $H\Psi^* = \mathcal{E}^*\Psi^* = \mathcal{E}^* e^{\mathcal{T}^*}\Psi_0$ by definition together with the commutativity of the cluster operators $\mathcal{T}^*$ and $\mathcal{S} = \sum_\nu \boldsymbol{s}_\nu \mathcal{X}_\nu$. □

Consider the setting of Corollary 2.4.1.1. Let us remark here that, thanks to Theorem 2.4.1, the eigenvalue $\mathcal{E}^*$ which appears in (2.22) coincides with the CC energy $\mathcal{E}^*_{\mathrm{CC}}$ generated by $\boldsymbol{t}^*$ through Equation (2.14). Therefore, when considering expressions of the form (2.22) involving the CC Fréchet derivative, we may refer to $\mathcal{E}^*$ as simply the coupled cluster energy associated with $\boldsymbol{t}^*$ without reference to the underlying eigenpair of the electronic Hamiltonian.

Corollary 2.4.1.2 (Local Lipschitz Continuity of Coupled Cluster Fréchet Derivative)**.**
Let the coupled cluster function $f\colon \mathbb{V}\to\mathbb{V}$ be defined through Definition 2.4.1, and for any $\boldsymbol{t}\in\mathbb{V}$ let $\mathrm{D}f(\boldsymbol{t})$ denote the Fréchet derivative of the coupled cluster function as defined through Equation (2.20)*. Then the mapping $\mathbb{V}\ni\boldsymbol{t}\mapsto \mathrm{D}f(\boldsymbol{t})\colon \mathbb{V}\to\mathbb{V}^*$ is Lipschitz continuous for bounded arguments, i.e., for any $\boldsymbol{t}\in\mathbb{V}$ and any any $\delta>0$, there exists a constant $L_{\boldsymbol{t}}(\delta)>0$ such that*

$$\sup_{\boldsymbol{t}\neq\boldsymbol{s}\in \mathrm{B}_\delta(\boldsymbol{t})} \frac{\|\mathrm{D}f(\boldsymbol{t}) - \mathrm{D}f(\boldsymbol{s})\|_{\mathbb{V}\to\mathbb{V}^*}}{\|\boldsymbol{t}-\boldsymbol{s}\|_{\mathbb{V}}} := L_{\boldsymbol{t}}(\delta) < \infty.$$

Corollary 2.4.1.2 follows immediately from the regularity of the coupled cluster function.

Having obtained an expression for the first Fréchet derivative $\mathrm{D}f(\boldsymbol{t})$, $\boldsymbol{t}\in\mathbb{V}$ of the coupled cluster function and studied some regularity properties of the mapping $\mathbb{V}\ni\boldsymbol{t}\mapsto \mathrm{D}f(\boldsymbol{t})$, the next step in our analysis will be to study the invertibility of the Fréchet derivative $\mathrm{D}f$ at any zero $\boldsymbol{t}^*\in\mathbb{V}$ of the coupled cluster function. In order to proceed with this analysis, let us first notice that thanks to the expression offered by Equation (2.22) in Corollary 2.4.1.1, the coupled cluster Fréchet derivative at any zero $\boldsymbol{t}^*\in\mathbb{V}$ can be described in terms of an operator acting on a subspace of the infinite-dimensional N-particle space $\widehat{\mathcal{H}}^1$. This observation motivates us to introduce the following operator acting on the space $\widetilde{\mathcal{V}} = \{\Psi_0\}^\perp \subset \widehat{\mathcal{H}}^1$ (recall Definition 2.3.4).

Definition 2.4.2 (Operator Induced by Coupled Cluster Fréchet Derivative at $\boldsymbol{t}^*$)**.**
Let the excitation index set $\mathcal{I}$ be defined through Definition 2.3.1, let the excitation operators $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{I}}$ be defined through Definition 2.3.2, let $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{I}}\in\mathbb{V}$ be any zero of the coupled cluster function defined through Definition 2.4.1, let $\mathcal{E}^*$ be the associated coupled cluster energy calculated through (2.14)*, and let the space $\widetilde{\mathcal{V}}\subset\widehat{\mathcal{H}}^1$ be defined as in Definition 2.3.4. We define the operator $\mathcal{A}(t^*)\colon \widetilde{\mathcal{V}}\to\widehat{\mathcal{H}}^{-1}$ as the mapping with the property that*

$$\forall\widetilde{\Psi}\in\widetilde{\mathcal{V}}\colon \qquad \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} := e^{-\mathcal{T}^*}\,(H-\mathcal{E}^*)\,e^{\mathcal{T}^*}\widetilde{\Psi} \qquad \text{where} \quad \mathcal{T}^* = \sum_{\mu\in\mathcal{I}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu. \tag{2.23}$$

Notation 2.4.2. *Let $\boldsymbol{t}^*\in\mathbb{V}$ be any zero of the coupled cluster function defined through Definition 2.4.1 and let $\mathcal{E}^*$ be the associated coupled cluster energy calculated through Equation* (2.14)*.*

- *We denote by* $\alpha_{\boldsymbol{t}^*} > 0$ *the constant defined as*

$$\alpha_{\boldsymbol{t}^*} := \|\mathcal{A}(\boldsymbol{t}^*)\|_{\widetilde{\mathcal{V}}\to\widetilde{\mathcal{V}}^*} := \sup_{0\neq\widetilde{\Phi}\in\widetilde{\mathcal{V}}} \sup_{0\neq\widetilde{\Psi}\in\widetilde{\mathcal{V}}} \frac{\langle\widetilde{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi}\rangle_{\widehat{\mathcal{H}}^{-1}\times\widehat{\mathcal{H}}^1}}{\|\widetilde{\Phi}\|_{\widehat{\mathcal{H}}^1}\|\widetilde{\Psi}\|_{\widehat{\mathcal{H}}^1}},$$

with the existence of $\alpha_{\boldsymbol{t}^*}$ *being guaranteed by Proposition 2.4.1.*

- *For any* $\boldsymbol{t} \in \mathbb{V}$ *we denote by* $L_{\boldsymbol{t}}\colon \mathbb{R}_+ \to \mathbb{R}_+$ *the so-called 'Lipschitz continuity function' as the mapping with the property that for all* $\delta > 0$ *it holds that*

$$L_{\boldsymbol{t}}(\delta) := \sup_{\boldsymbol{t}\neq\boldsymbol{s}\in\mathrm{B}_\delta(\boldsymbol{t})} \frac{\|\mathrm{D}f(\boldsymbol{t}) - \mathrm{D}f(\boldsymbol{s})\|_{\mathbb{V}\to\mathbb{V}^*}}{\|\boldsymbol{t}-\boldsymbol{s}\|_{\mathbb{V}}},$$

with the existence of the function $L_{\boldsymbol{t}}$ *being guaranteed by Corollary 2.4.1.2.*

- *We denote by* $\mathcal{A}(\boldsymbol{t}^*)^\dagger\colon \widetilde{\mathcal{V}} \to \widehat{\mathcal{H}}^{-1}$ *the mapping with the property that for all* $\widetilde{\Psi} \in \widetilde{\mathcal{V}}$ *it holds that*

$$\mathcal{A}(\boldsymbol{t}^*)^\dagger\widetilde{\Psi} := e^{(\mathcal{T}^*)^\dagger}\,(H - \mathcal{E}^*)\,e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}, \tag{2.24}$$

and we emphasise that for all $\widetilde{\Psi}, \widetilde{\Phi} \in \widetilde{\mathcal{V}}$ *it holds that*

$$\langle\widetilde{\Phi}, \mathcal{A}(\boldsymbol{t}^*)^\dagger\widetilde{\Psi}\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} = \langle\mathcal{A}(\boldsymbol{t}^*)\widetilde{\Phi}, \widetilde{\Psi}\rangle_{\widehat{\mathcal{H}}^{-1}\times\widehat{\mathcal{H}}^1},$$

so that in particular

$$\|\mathcal{A}(\boldsymbol{t}^*)^\dagger\|_{\widetilde{\mathcal{V}}\to\widetilde{\mathcal{V}}^*} = \|\mathcal{A}(\boldsymbol{t}^*)\|_{\widetilde{\mathcal{V}}\to\widetilde{\mathcal{V}}^*} = \alpha_{\boldsymbol{t}^*}.$$

Consider now Definition 2.4.2 of the bounded linear operator $\mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}} \to \widehat{\mathcal{H}}^{-1}$ for an arbitrary zero $\boldsymbol{t}^* \in \mathbb{V}$ of the coupled cluster function. Since the coefficient space $\mathbb{V}$ inherits its inner product from the inner product on $\widehat{\mathcal{H}}^1$, it immediately follows that

$$\mathrm{D}f(\boldsymbol{t}^*)\colon \mathbb{V}\to\mathbb{V}^* \quad \text{is an isomorphism} \qquad \Longleftrightarrow \qquad \mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}}\to\widetilde{\mathcal{V}}^* \quad \text{is an isomorphism.}$$

We claim that the mapping $\mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ can indeed be shown to be an isomorphism *provided* that the zero $\boldsymbol{t}^*$ is generated by an *intermediately normalisable eigenfunction* $\Psi^* \in \widehat{\mathcal{H}}^1$ of the electronic Hamiltonian that corresponds to a *simple and isolated eigenvalue*. The proof of this claim, which is the subject of the next theorem, is based on classical functional analysis arguments, and will proceed in the following steps: Assuming that the zero $\boldsymbol{t}^*$ is generated by a non-degenerate, intermediately normalisable eigenfunction of the electronic Hamiltonian:

1. We will first show that $\mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is injective. As a consequence of the Hahn-Banach theorem, we will deduce that the adjoint operator $\mathcal{A}(\boldsymbol{t}^*)^\dagger\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ has *dense* range.

2. Next, we will show that the operator $\mathcal{A}(\boldsymbol{t}^*)^\dagger\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is bounded below. This will imply that $\mathcal{A}(\boldsymbol{t}^*)^\dagger\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is injective, and has *closed* range.

Combining the above two steps, will allow us to deduce that the adjoint operator $\mathcal{A}(\boldsymbol{t}^*)^\dagger\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is an isomorphism, and therefore so too is the operator $\mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$. Let us emphasise here that rather than attacking directly the operator $\mathcal{A}(\boldsymbol{t}^*)$ induced by the Fréchet derivative of the coupled cluster function at $\boldsymbol{t}^* \in \mathbb{V}$, we are choosing to analyse its adjoint. This choice is

motivated by practical reasons: there is a technical difficulty in proving directly the invertibility of $\mathcal{A}(\boldsymbol{t}^*)$ which is avoided if we study instead $\mathcal{A}(\boldsymbol{t}^*)^\dagger$.

Theorem 2.4.3 (Invertibility of Operator Induced by Coupled Cluster Fréchet Derivative at $\boldsymbol{t}^*$)**.**

Let $\boldsymbol{t}^* = \{\boldsymbol{t}^*_\mu\}_{\mu\in\mathcal{G}} \in \mathbb{V}$ be associated with a non-degenerate, intermediately normalisable eigenpair $(\mathcal{E}^*, \Psi^*) \in \mathbb{R}\times\widehat{\mathcal{H}}^1$ of the electronic Hamiltonian $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ defined through Equation (2.2), i.e.,

$$H\Psi^* = \mathcal{E}^*\Psi^*, \qquad \text{with } \mathcal{E}^* \text{ simple, isolated} \qquad \text{and} \qquad \Psi^* = e^{\mathcal{T}^*}\Psi_0 \quad \text{where} \quad \mathcal{T}^* = \sum_{\mu\in\mathcal{G}} \boldsymbol{t}^*_\mu \mathcal{X}_\mu.$$

Then the operator $\mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ defined through Definition 2.4.2 is an isomorphism.

Proof. The proof follows the aforementioned two steps. We begin with the injectivity of $\mathcal{A}(\boldsymbol{t}^*)$.

Step 1: $\mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is injective.

Suppose there exists $0 \neq \widetilde{\Psi} \in \widetilde{\mathcal{V}}$ such that $\mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \equiv 0$ in $\widetilde{\mathcal{V}}^*$, i.e., for all $\widetilde{\Phi} \in \widetilde{\mathcal{V}}$ it holds that

$$\left\langle \widetilde{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} = 0. \tag{2.25}$$

As a first step, we claim that from Equation (2.25) it must follow that $\mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \equiv 0$ in $\widehat{\mathcal{H}}^{-1}$. Recalling the complementary decomposition of $\widehat{\mathcal{H}}^1$ given by Definition 2.3.5, we see that it suffices to prove that

$$\left\langle \Psi_0, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} = 0. \tag{2.26}$$

Consider now the element $\widehat{\Phi} = e^{(\mathcal{T}^*)^\dagger} e^{\mathcal{T}^*}\Psi_0 \in \widehat{\mathcal{H}}^1$ and recall that we denote by $\mathbb{P}_0\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ the $\widehat{\mathcal{L}}^2$-orthogonal projection operator onto $\mathrm{span}\{\Psi_0\}$ defined through Definition 2.3.5 and we have defined $\mathbb{P}_0^\perp := \mathbb{I} - \mathbb{P}_0$. Clearly, we have that $\mathbb{P}_0\widehat{\Phi} \neq 0$ since

$$(\widehat{\Phi}, \Psi_0)_{\widehat{\mathcal{L}}^2} = \left(e^{(\mathcal{T}^*)^\dagger} e^{\mathcal{T}^*}\Psi_0, \Psi_0\right)_{\widehat{\mathcal{L}}^2} = \left(e^{\mathcal{T}^*}\Psi_0, e^{\mathcal{T}^*}\Psi_0\right)_{\widehat{\mathcal{L}}^2} = \|\Psi^*\|^2_{\widehat{\mathcal{L}}^2} =: \widehat{d}_0 \neq 0. \tag{2.27}$$

Since $\Psi^* = e^{\mathcal{T}^*}\Psi_0 \in \widehat{\mathcal{H}}^1$ is by definition an eigenfunction of the electronic Hamiltonian with associated eigenvalue $\mathcal{E}^*$, a direct calculation also reveals that

$$\begin{aligned}
\left\langle \widehat{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} &= \left\langle e^{(\mathcal{T}^*)^\dagger} e^{\mathcal{T}^*}\Psi_0, e^{-\mathcal{T}^*}\left(H - \mathcal{E}^*\right) e^{\mathcal{T}^*}\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \\
&= \left\langle e^{\mathcal{T}^*}\Psi_0, \left(H - \mathcal{E}^*\right) e^{\mathcal{T}^*}\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \\
&= 0.
\end{aligned} \tag{2.28}$$

On the other hand, we also have

$$\left\langle \widehat{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = \left\langle \mathbb{P}_0\widehat{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} + \left\langle \mathbb{P}_0^{\perp}\widehat{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = \left\langle \mathbb{P}_0\widehat{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}, \tag{2.29}$$

where the second equality is due to the fact that $\mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \equiv 0$ in $\widetilde{\mathcal{V}}^*$ by assumption.

Combining therefore Equations (2.27)-(2.29), we deduce that

$$0 = \left\langle \widehat{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = \left\langle \mathbb{P}_0\widehat{\Phi}, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = \widehat{d}_0 \left\langle \Psi_0, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}.$$

Since $\widehat{d}_0 \neq 0$, we immediately deduce that Equation (2.26) holds and therefore $\mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \equiv 0$ in $\widehat{\mathcal{H}}^{-1}$ as claimed.

Since $e^{-(\mathcal{T}^*)^\dagger} : \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ is a bijection, we next deduce that for all $\Phi \in \widehat{\mathcal{H}}^1$ it holds that

$$0 = \left\langle e^{(\mathcal{T}^*)^\dagger}\Phi, \mathcal{A}(\boldsymbol{t}^*)\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} = \left\langle \Phi, (H - \mathcal{E}^*)\, e^{\mathcal{T}^*}\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}.$$

The simplicity of the eigenvalue $\mathcal{E}^*$ now implies that we must have

$$e^{\mathcal{T}^*}\widetilde{\Psi} \in \quad \mathrm{span}\{\Psi^*\}.$$

Using again the fact that $\Psi^* = e^{\mathcal{T}^*}\Psi_0$ and that $e^{\mathcal{T}^*} : \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ is a bijection, we obtain the existence of some constant $\widetilde{d}_0 \in \mathbb{R}$ such that

$$\widetilde{\Psi} = \widetilde{d}_0\Psi_0.$$

Recall however that $\widetilde{\Psi} \in \widetilde{\mathcal{V}} = \{\Psi_0\}^{\perp}$ by assumption, and therefore we must have $\widetilde{d}_0 = 0$ and thus $\widetilde{\Psi} = 0$. This completes the proof of the first step.

Step 2: $\mathcal{A}(\boldsymbol{t}^*)^\dagger : \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is bounded below.

Let $\widetilde{\Psi} \in \widetilde{\mathcal{V}}$ be arbitrary. For any $\Psi_\perp^* \in \{\Psi^*\}^{\perp} \subset \widehat{\mathcal{H}}^1$, i.e., any wave-function $\Psi_\perp^*$ that is $\widehat{\mathcal{L}}^2$-orthogonal to the eigenfunction Ψ^* with associated eigenvalue $\mathcal{E}^*$, we define the function

$$\widetilde{\Phi}_{\Psi_\perp} := \mathbb{P}_0^{\perp} e^{-\mathcal{T}^*}\Psi_\perp^* \in \widetilde{\mathcal{V}},$$

It is straightforward to observe that for all such $\widetilde{\Phi}_{\Psi^\perp}$, it holds that

$$\begin{aligned}\left| \left\langle \widetilde{\Phi}_{\Psi_\perp}, \mathcal{A}(\boldsymbol{t}^*)^\dagger\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} \right| = \Big| & \underbrace{\left\langle e^{-\mathcal{T}^*}\Psi_\perp^*, e^{(\mathcal{T}^*)^\dagger}\,(H - \mathcal{E}^*)\, e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{I})} \\ & - \underbrace{\left\langle \mathbb{P}_0 e^{-\mathcal{T}^*}\Psi_\perp^*, e^{(\mathcal{T}^*)^\dagger}\,(H - \mathcal{E}^*)\, e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{II})} \Big|.\end{aligned} \tag{2.30}$$

We claim that the term (II) is identically zero for any choice of $\Psi_\perp^*$. To this end, observe that

$$\mathbb{P}_0 e^{-\mathcal{T}^*}\Psi_\perp^* = \left(e^{-\mathcal{T}^*}\Psi_\perp^*, \Psi_0\right)_{\widehat{\mathcal{L}}^2}\Psi_0 = (\Psi_\perp^*, \Psi_0)_{\widehat{\mathcal{L}}^2}\,\Psi_0 = \mathbb{P}_0\Psi_\perp^*.$$

We therefore deduce that

$$\begin{aligned}(\mathrm{II}) &= -\left\langle \mathbb{P}_0\Psi_\perp^*, e^{(\mathcal{T}^*)^\dagger}\,(H-\mathcal{E}^*)\,e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \\ &= -\,(\Psi_0, \Psi_\perp^*)_{\widehat{\mathcal{L}}^2}\,\left\langle \Psi_0, e^{(\mathcal{T}^*)^\dagger}\,(H-\mathcal{E}^*)\,e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}},\end{aligned}$$

where we have used the fact that $\mathbb{P}_0\Psi_\perp^* = (\Psi_0, \Psi_\perp^*)_{\widehat{\mathcal{L}}^2}\,\Psi_0$.

Notice however that the second term in the product above satisfies

$$\left\langle \Psi_0, e^{(\mathcal{T}^*)^\dagger}\,(H-\mathcal{E}^*)\,e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} = \left\langle (H-\mathcal{E}^*)\,e^{\mathcal{T}^*}\Psi_0, e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^{-1}\times\widehat{\mathcal{H}}^{1}} = 0, \tag{2.31}$$

where the last step follows from the fact that $He^{\mathcal{T}^*}\Psi_0 = H\Psi^* = \mathcal{E}^*\Psi^*$ by assumption. Thus, the term (II) is identically zero for any choice of $\Psi_\perp^* \in \{\Psi^*\}^\perp \subset \widehat{\mathcal{H}}^1$ as claimed, and we need only estimate the term (I).

An easy simplification reveals that

$$(\mathrm{I}) = \left\langle \Psi_\perp^*, (\mathrm{H}-\mathcal{E}^*)\,\mathrm{e}^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}. \tag{2.32}$$

Thanks to the ellipticity of the electronic Hamiltonian $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ and the simplicity of the eigenvalue $\mathcal{E}^*$, it is easy to deduce that the shifted Hamiltonian $H-\mathcal{E}^*\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ satisfies an inf-sup condition on $\{\Psi^*\}^\perp \subset \widehat{\mathcal{H}}^1$ (see also Remark 2.4.3 for a detailed argument). In order to make use of this result and bound the term (I), we need only show that $\Psi_\perp^*$ and $e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}$ are both elements of $\{\Psi^*\}^\perp$. The former inclusion is true by definition of $\Psi_\perp^*$ and as for latter, we see that

$$\left(\Psi^*, e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right)_{\widehat{\mathcal{L}}^2} = \left(e^{\mathcal{T}^*}\Psi_0, e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right)_{\widehat{\mathcal{L}}^2} = \left(\Psi_0, \widetilde{\Psi}\right)_{\widehat{\mathcal{L}}^2} = 0,$$

where we have used the fact that $\widetilde{\Psi} \in \widetilde{\mathcal{V}} = \{\Psi_0\}^\perp$ by definition.

We can therefore deduce from Equation (2.32) that

$$\sup_{\Psi_\perp^*\in\{\Psi^*\}^\perp} \frac{\left|\left\langle \Psi_\perp^*, (H-\mathcal{E}^*)\,e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{\|\Psi_\perp^*\|_{\widehat{\mathcal{H}}^1}} \geqslant \gamma\left\|e^{-(\mathcal{T}^*)^\dagger}\widetilde{\Psi}\right\|_{\widehat{\mathcal{H}}^1}, \tag{2.33}$$

where $\gamma > 0$ is the inf-sup constant of the shifted Hamiltonian $H-\mathcal{E}^*$ on $\{\Psi^*\}^\perp \subset \widehat{\mathcal{H}}^1$.

Recalling now that $\widetilde{\Psi} \in \widetilde{\mathcal{V}}$ was arbitrary and combining the estimates (2.31)-(2.33) with

Equation (2.30) we obtain that for all $\widetilde{\Psi} \in \widetilde{\mathcal{V}}$ it holds that

$$\begin{aligned}
\|\mathcal{A}(\boldsymbol{t}^*)^\dagger \widetilde{\Psi}\|_{\widetilde{\mathcal{V}}^*} &= \sup_{0 \neq \widetilde{\Phi} \in \widetilde{\mathcal{V}}} \frac{\left| \left\langle \widetilde{\Phi}, \mathcal{A}(\boldsymbol{t}^*)^\dagger \widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} \right|}{\|\widetilde{\Phi}\|_{\widehat{\mathcal{H}}^1}} \geqslant \sup_{0 \neq \Psi_\perp^* \in \{\Psi^*\}^\perp} \frac{\left| \left\langle \widetilde{\Phi}_{\Psi_\perp}, \mathcal{A}(\boldsymbol{t}^*)^\dagger \widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} \right|}{\|\widetilde{\Phi}_{\Psi_\perp}\|_{\widehat{\mathcal{H}}^1}} \\
&= \sup_{0 \neq \Psi_\perp^* \in \{\Psi^*\}^\perp} \frac{\left| \left\langle \Psi_\perp^*, (H - \mathcal{E}^*)\, e^{-(\mathcal{T}^*)^\dagger} \widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} \right|}{\|\mathbb{P}_0^\perp e^{-\mathcal{T}(\boldsymbol{t}^*)} \Psi_\perp^*\|_{\widehat{\mathcal{H}}^1}} \\
&\geqslant \frac{1}{\|\mathbb{P}_0^\perp e^{-\mathcal{T}^*}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1}} \sup_{0 \neq \Psi_\perp^* \in \{\Psi^*\}^\perp} \frac{\left| \left\langle \Psi_\perp^*, (H - \mathcal{E}^*)\, e^{-(\mathcal{T}^*)^\dagger} \widetilde{\Psi} \right\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} \right|}{\|\Psi_\perp^*\|_{\widehat{\mathcal{H}}^1}} \\
&\geqslant \frac{\gamma}{\|\mathbb{P}_0^\perp e^{-\mathcal{T}^*}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1}} \left\| e^{-(\mathcal{T}^*)^\dagger} \widetilde{\Psi} \right\|_{\widehat{\mathcal{H}}^1} \\
&\geqslant \frac{\gamma}{\|\mathbb{P}_0^\perp e^{-\mathcal{T}^*}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1} \|e^{(\mathcal{T}^*)^\dagger}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1}} \|\widetilde{\Psi}\|_{\widehat{\mathcal{H}}^1},
\end{aligned}$$

where the final step follows from the fact that $e^{-(\mathcal{T}^*)^\dagger} : \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ is a bijection. Defining now the constant $\Theta \in (0, \infty)$ as

$$\Theta := \|e^{(\mathcal{T}^*)^\dagger}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1} \|\mathbb{P}_0^\perp e^{-\mathcal{T}^*}\|_{\widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1}, \tag{2.34}$$

and recalling that $\widetilde{\Psi} \in \widetilde{\mathcal{V}}$ was arbitrary, we deduce that

$$\forall \widetilde{\Psi} \in \widetilde{\mathcal{V}}: \quad \|\mathcal{A}(\boldsymbol{t}^*)^\dagger \widetilde{\Psi}\|_{\widetilde{\mathcal{V}}^*} \geqslant \frac{\gamma}{\Theta} \|\widetilde{\Psi}\|_{\widehat{\mathcal{H}}^1},$$

which completes the proof of the second step.

Combining the conclusions of **Step 1** and **Step 2** we deduce that the adjoint operator $\mathcal{A}(\boldsymbol{t}^*)^\dagger : \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is an isomorphism, and from this it follows that the operator $\mathcal{A}(\boldsymbol{t}^*) : \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ is also an isomorphism. □

Equipped with Theorem 2.4.3 and recalling the discussion following Notation 2.4.2, we immediately obtain the desired invertibility result for the coupled cluster Fréchet derivative at any zero $\boldsymbol{t}^* \in \mathbb{V}$ of the coupled cluster function that is associated with a non-degenerate, intermediately normalised eigenfunction $\Psi^* \in \widehat{\mathcal{H}}^1$ of the electronic Hamiltonian.

Corollary 2.4.3.1 (Invertibility of the Coupled Cluster Fréchet Derivative at $\boldsymbol{t}^*$)**.**

Let the coupled cluster function $f : \mathbb{V} \to \mathbb{V}^*$ be defined through Definition 2.4.1, for any $\boldsymbol{t} \in \mathbb{V}$ let $\mathrm{D}f(\boldsymbol{t})$ denote the Fréchet derivative of the coupled cluster function as defined through Equation (2.20), let $\boldsymbol{t}^* \in \mathbb{V}$ denote a zero of the coupled cluster function corresponding to an intermediately normalised eigenfunction $\Psi^* \in \widehat{\mathcal{H}}^1$ of the electronic Hamiltonian $H : \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ with simple, isolated eigenvalue $\mathcal{E}^*$, let $\gamma > 0$ denote the inf-sup constant of the shifted Hamiltonian $H - \mathcal{E}^*$ on $\{\Psi^*\}^\perp \subset \widehat{\mathcal{H}}^1$, and let $\Theta > 0$ denote the constant defined through Equation (2.34). Then $\mathrm{D}f(\boldsymbol{t}^*) : \mathbb{V} \to \mathbb{V}^*$ is an isomorphism and it holds that

$$\|\mathrm{D}f(\boldsymbol{t}^*)^{-1}\|_{\mathbb{V}^* \to \mathbb{V}} \leqslant \frac{\Theta}{\gamma}.$$

Having completed our study of the coupled cluster Fréchet derivative, we are now finally ready to state the main result of this section, namely the local well-posedness of the single reference coupled cluster equations. As mentioned at the beginning of this section, we will do so by appealing to a classical result from non-linear numerical analysis.

Theorem 2.4.4 (Local Uniqueness of the Coupled Cluster Solution $\boldsymbol{t}^*$)**.**
Let the coupled cluster function $f\colon \mathbb{V} \to \mathbb{V}^*$ be defined through Definition 2.4.1, let $\boldsymbol{t}^* \in \mathbb{V}$ denote a zero of the coupled cluster function corresponding to an intermediately normalised eigenfunction $\Psi^* \in \widehat{\mathcal{H}}^1$ of the electronic Hamiltonian $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ with simple, isolated eigenvalue $\mathcal{E}^*$, let $\gamma > 0$ denote the inf-sup constant of the shifted Hamiltonian $H - \mathcal{E}^*$ on $\{\Psi^*\}^\perp \subset \widehat{\mathcal{H}}^1$, let $\Theta > 0$ denote the constant defined through Equation (2.34), let the continuity constant $\alpha_{\boldsymbol{t}^*} > 0$ and the Lipschitz continuity function $L_{\boldsymbol{t}^*}\colon \mathbb{R}_+ \to \mathbb{R}_+$ be defined according to Notation 2.4.2, and define the constant

$$\mathrm{T} := \min_{\delta > 0} \left\{ \delta,\ \frac{\gamma}{L_{\boldsymbol{t}^*}(\delta)\Theta},\ 2\frac{\alpha_{\boldsymbol{t}^*}}{L_{\boldsymbol{t}^*}(\delta)} \right\}.$$

Then $f\big(\mathrm{B}_{\mathrm{T}}(\boldsymbol{t}^*)\big)$ is an open subset $\mathbb{V}^*$, the restriction of f to $\mathrm{B}_{\mathrm{T}}(\boldsymbol{t}^*)$ is a diffeomorphism *and for all $\boldsymbol{s} \in \mathrm{B}_{\mathrm{T}}(\boldsymbol{t}^*)$ we have the error estimate*

$$\frac{1}{2}\frac{1}{\alpha_{\boldsymbol{t}^*}}\|f(\boldsymbol{s})\|_{\mathbb{V}^*} \leqslant \|\boldsymbol{t}^* - \boldsymbol{s}\|_{\mathbb{V}} \leqslant 2\frac{\Theta}{\gamma}\|f(\boldsymbol{s})\|_{\mathbb{V}^*}. \tag{2.35}$$

In particular, $\boldsymbol{t}^*$ is the unique solution of the continuous coupled cluster equations (2.13) in the open ball $\mathrm{B}_{\mathrm{T}}(\boldsymbol{t}^*)$.

Proof. The fact that the image under f of the open ball $\mathrm{B}_{\mathrm{T}}(\boldsymbol{t}^*)$ is itself open and that f is a local diffeomorphism is a direct consequence of the inverse function theorem for Banach spaces (see, e.g., [66, Chapter 9]) while the error estimate is a direct application of [96, Proposition 2.1]. The fact that the assumptions of both results are indeed fulfilled by the coupled cluster function f is a consequence of Proposition 2.4.1 and Corollaries 2.4.1.2 and 2.4.3.1. □

Next, let us comment on the constants that appear in the error estimate offered by Theorem 2.4.4.

Remark 2.4.3 (Interpretation of the Constants Appearing in Error Estimate (2.35))**.**
Consider the setting of Theorem 2.4.4. From the point of view of a posteriori error quantification, it is important to gain a better understanding of the constants $\gamma > 0$ and $\Theta > 0$.

Let us recall that the $\gamma > 0$ is the inf-sup constant of the shifted Hamiltonian $H - \mathcal{E}^*$ on $\{\Psi^*\}^\perp \subset \widehat{\mathcal{H}}^1$. A crude lower bound for this constant can be obtained through the following procedure.

We begin by noting that the shifted Hamiltonian $H - \mathcal{E}^*_{\mathrm{GS}} + 1$ defines a coercive operator on $\widehat{\mathcal{H}}^1$. Since the electronic Hamiltonian is additionally self-adjoint, we can introduce a new norm on $\widehat{\mathcal{H}}^1$ by setting

$$\forall \Phi \in \widehat{\mathcal{H}}^1\colon \qquad |||\Phi|||^2_{\widehat{\mathcal{H}}^1} := \langle \Phi, (H - \mathcal{E}^*_{\mathrm{GS}} + 1)\,\Phi\rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}},$$

and it is clear that this new norm is equivalent to the canonical $\|\cdot\|_{\widehat{\mathcal{H}}^1}$ norm, i.e.,

$$\exists c_{\mathrm{equiv}} > 1 \ \text{ such that } \ \forall \Phi \in \widehat{\mathcal{H}}^1\colon \quad \frac{1}{c_{\mathrm{equiv}}}|||\Phi|||_{\widehat{\mathcal{H}}^1} \leqslant \|\Phi\|_{\widehat{\mathcal{H}}^1} \leqslant c_{\mathrm{equiv}}|||\Phi|||_{\widehat{\mathcal{H}}^1}.$$

In particular, the ellipticity of the electronic Hamiltonian given by Inequality (2.5) *also holds with respect to the new* $|||\cdot|||_{\widehat{\mathcal{H}}^1}$ *norm and we have*

$$\forall \Phi \in \widehat{\mathcal{H}}^1\colon \quad \langle \Phi, (H-\mathcal{E}^*)\,\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \geqslant \frac{1}{4c_{\mathrm{equiv}}}|||\Phi|||^2_{\widehat{\mathcal{H}}^1} - \left(9QZ^2 - \mathcal{E}^* - \frac{1}{4}\right)\|\Phi\|^2_{\widehat{\mathcal{L}}^2}. \tag{2.36}$$

Moreover, the norm $|||\cdot|||_{\widehat{\mathcal{H}}^1}$ *also induces a new dual norm* $|||\cdot|||_{\widehat{\mathcal{H}}^{-1}}$ *on the space* $\widehat{\mathcal{H}}^{-1}$*, and this new norm is also equivalent to the canonical dual norm* $\|\cdot\|_{\widehat{\mathcal{H}}^{-1}}$*.*

Next, we claim that for any $\Phi \in \{\Psi^*\}^\perp \subset \widehat{\mathcal{H}}^1$ *there exists* $\Phi_{\mathrm{flip}} \in \{\Psi^*\}^\perp$ *such that*

$$\langle \Phi_{\mathrm{flip}}, (H-\mathcal{E}^*)\,\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \geqslant \Lambda^*\|\Phi\|^2_{\widehat{\mathcal{L}}^2}, \quad \text{where} \quad \Lambda^* := \inf_{\substack{\lambda^*\in\sigma(H)\\ \lambda^*\neq\mathcal{E}^*}} |\lambda^*-\mathcal{E}^*| > 0 \quad \text{is the spectral gap at } \mathcal{E}^*. \tag{2.37}$$

To see this, assume that $(\mathcal{E}^*, \Psi^*)$ *is the* J^{th} *eigenpair of the electronic Hamiltonian, ordered non-decreasingly and counting multiplicity. Then we can write any* $\Phi \in \{\Psi^*\}^\perp$ *in the form*

$$\Phi = \sum_{\ell=1}^{J+1} \mathbb{P}_\ell \Phi + \Phi^\perp,$$

where each $\mathbb{P}_\ell\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$ *denotes the* $\widehat{\mathcal{L}}^2$*-orthogonal projector onto the span of the* ℓ^{th} *eigenfunction,* $\Phi^\perp := \Phi - \sum_{\ell=1}^{J+1}\mathbb{P}_j\Phi$*, and we emphasise that* $\mathbb{P}_J\Phi = 0$ *since* $\Phi \in \{\Psi^*\}^\perp$*. Consequently, for any* $\Phi \in \{\Psi^*\}^\perp$*, we may define* $\Phi_{\mathrm{flip}} \in \{\Psi^*\}^\perp$ *as*

$$\Phi_{\mathrm{flip}} := -\sum_{\ell=1}^{J-1} \mathbb{P}_\ell\Phi + \mathbb{P}_{J+1}\Phi + \Phi^\perp,$$

and a direct calculation shows that

$$\langle \Phi_{\mathrm{flip}}, (H-\mathcal{E}^*)\,\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \geqslant \min\{\mathcal{E}^*-\mathcal{E}_{J-1}, \mathcal{E}_{J+1}-\mathcal{E}^*\}\,\|\Phi\|^2_{\widehat{\mathcal{L}}^2} =: \Lambda^*\|\Phi\|^2_{\widehat{\mathcal{L}}^2},$$

where we have used $\mathcal{E}_{J-1}, \mathcal{E}_{J+1}$ *to denote the* $J-1$ *and* $J+1$ *eigenvalues of the electronic Hamiltonian. The claim now readily follows. Additionally, it is readily verified that for any* $\Phi \in \{\Psi^*\}^\perp$ *with* Φ_{flip} *constructed according to the above procedure, it holds that* $|||\Phi_{\mathrm{flip}}|||_{\widehat{\mathcal{H}}^1} = |||\Phi|||_{\widehat{\mathcal{H}}^1}$*.*

Defining now the constant $q := \dfrac{\Lambda^*}{\Lambda^* + \left(9QZ^2 - \mathcal{E}^* - \frac{1}{4}\right)} \in (0,1)$ *and combining the Estimates* (2.36) *and* (2.37)*, we deduce that for all* $\Phi \in \{\Psi^*\}^\perp$ *it holds that*

$$\begin{aligned}
\sup_{0\neq\Psi\in\{\Psi^*\}^\perp} \frac{\left|\langle\Psi,(H-\mathcal{E}^*)\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Psi|||_{\widehat{\mathcal{H}}^1}} &= q \sup_{0\neq\Psi\in\{\Psi^*\}^\perp} \frac{\left|\langle\Psi,(H-\mathcal{E}^*)\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Psi|||_{\widehat{\mathcal{H}}^1}} \\
&\quad + (1-q) \sup_{0\neq\Psi\in\{\Psi^*\}^\perp} \frac{\left|\langle\Psi,(H-\mathcal{E}^*)\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Psi|||_{\widehat{\mathcal{H}}^1}} \\
&\geqslant q\frac{\left|\langle\Phi,(H-\mathcal{E}^*)\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Phi|||_{\widehat{\mathcal{H}}^1}} + (1-q)\frac{\left|\langle\Phi_{\mathrm{flip}},(H-\mathcal{E}^*)\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Phi_{\mathrm{flip}}|||_{\widehat{\mathcal{H}}^1}} \\
&\geqslant q\frac{1}{|||\Phi|||_{\widehat{\mathcal{H}}^1}}\left(\frac{1}{4\ c_{\mathrm{equiv}}}|||\Phi|||^2_{\widehat{\mathcal{H}}^1} - \left(9QZ^2-\mathcal{E}^*-\frac{1}{4}\right)\|\Phi\|^2_{\widehat{\mathcal{L}}^2}\right) \\
&\quad + (1-q)\frac{1}{|||\Phi|||_{\widehat{\mathcal{H}}^1}}\Lambda^*\|\Phi\|^2_{\widehat{\mathcal{L}}^2} \\
&= q\frac{1}{4\ c_{\mathrm{equiv}}}|||\Phi|||_{\widehat{\mathcal{H}}^1},
\end{aligned}$$

where the cancellations in the last step occurs due to the definition of $q \in (0,1)$.

Recalling now the definition of the constant q*, we see that the inf-sup constant* γ *is lower bounded by*

$$\gamma \geqslant \frac{\Lambda^*}{4\ c_{\mathrm{equiv}}\left(\Lambda^* + 9QZ^2 - \mathcal{E}^* - \frac{1}{4}\right)}. \tag{2.38}$$

Two important comments are now in order. First, we expect the lower bound (2.38) *to be rather coarse because of the appearence of the norm equivalence constant* c_{equiv}*. Note also that the* $\widehat{\mathcal{H}}^1$*-norm associated with this equivalance constant is given by*

$$\forall\Phi\in\widehat{\mathcal{H}}^1: \qquad |||\Phi|||^2_{\widehat{\mathcal{H}}^1} = \langle\Phi,(H-\mathcal{E}^*_{\mathrm{GS}}+1)\,\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}},$$

i.e., as the norm induced by the shifted Hamiltonian $H-\mathcal{E}^*_{\mathrm{GS}}+1$*, and there is* a priori *no reason that a better equivalence constant cannot be obtained for a differently shifted Hamiltonian, i.e., for the operator* $H-\mathcal{E}^*_{\mathrm{GS}}+\iota$ *with* $\iota>0$ *arbitrary.*

Second, we observe that if Λ^**, i.e., the spectral gap at* $\mathcal{E}^*$*, approaches zero, then the lower bound* (2.38) *that we have derived also approches zero. In fact, the same is true for the inf-sup constant* γ*, i.e.,* $\Lambda^*\to 0$ *implies that* $\gamma\to 0$*. To see this, assume for simplicity that* $\Lambda^* = \mathcal{E}^* - \widetilde{\mathcal{E}}$ *with* $\widetilde{\mathcal{E}}$ *denoting the eigenvalue associated with some eigenfunction* $\widetilde{\Psi}\neq\Psi^*\in\widehat{\mathcal{H}}^1$ *of the electronic*

Hamiltonian. It then follows that

$$\gamma = \inf_{0\neq\Phi\in\{\Psi^*\}^\perp}\sup_{0\neq\Psi\in\{\Psi^*\}^\perp}\frac{\left|\langle\Psi,(H-\mathcal{E}^*)\Phi\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Psi|||_{\widehat{\mathcal{H}}^1}|||\Phi|||_{\widehat{\mathcal{H}}^1}} \leqslant \sup_{0\neq\Psi\in\{\Psi^*\}^\perp}\frac{\left|\left\langle\Psi,(H-\mathcal{E}^*)\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Psi|||_{\widehat{\mathcal{H}}^1}|||\widetilde{\Psi}|||_{\widehat{\mathcal{H}}^1}}$$
$$=\Lambda^*\sup_{0\neq\Psi\in\{\Psi^*\}^\perp}\frac{\left|\left\langle\Psi,\widetilde{\Psi}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\right|}{|||\Psi|||_{\widehat{\mathcal{H}}^1}|||\widetilde{\Psi}|||_{\widehat{\mathcal{H}}^1}},$$

from which we deduce that $\Lambda^*\to 0$ indeed implies $\gamma\to 0$.

An important consequence of this observation is that the residual-based CC error estimate (2.35) *that we have derived in this work will degrade as the spectral gap degrades. In particular, it will not hold for gapless systems for which a more elaborate theory must be developed. A possible starting point for such a theory could be the Lyapunov–Schmidt construction (see, e.g., [12, Chapter V]) which is used in non-linear numerical analysis to study problems that cannot be analysed using the inverse function theorem. Of course, the applicability of this approach to the coupled cluster equations is an open question.*

Coming now to the constant Θ, we see that it is simply the product of two operator norms involving the exponential cluster operator and its adjoint. Thanks to the continuity of the mapping $\mathbb{V}\ni\boldsymbol{t}\mapsto e^{-\mathcal{T}(\boldsymbol{t})}\colon\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1$, (and its adjoint) we deduce that these operator norms will be large when $\|\boldsymbol{t}\|_{\mathbb{V}}$ is large, and therefore the residual-based CC error estimate (2.35) *is expected to degrade if $\|\boldsymbol{t}\|_{\mathbb{V}}$ is large.*

We conclude this section by emphasising, in particular, that if the ground state energy of the electronic Hamiltonian $H\colon\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^{-1}$ is simple, and the chosen reference determinant Ψ_0 is not orthogonal to the corresponding ground state wave-function, then the continuous coupled cluster equations (2.17) are locally well-posed, and we have access to the residual-based error estimates given by Theorem 2.4.4.

2.5 Well-posedness of the full coupled cluster equations in a finite basis

Having understood the local well-posedness of the continuous coupled cluster function, the next step in our analysis is to study the discrete coupled cluster equations (2.15). Unfortunately, obtaining a local well-posedness result for an arbitrary choice of excitation subset or Q-particle basis set is a highly non-trivial exercise. Indeed, as the subsequent exposition will show (see Lemma 2.5.2 and Theorem 2.5.5 below), our discrete local well-posedness analysis depends on being able to demonstrate that certain discrete inf-sup conditions hold and the establishment of these conditions for arbitrary discretisations is not obvious. For the purpose of this thesis therefore, we will limit ourselves to an analysis of the so-called Full-Coupled Cluster equations in a finite basis. The extension of our analysis to more general discretisations (the so-called *truncated* CC equations [52, Chapter 13]) will be addressed in a forthcoming contribution.

Throughout this section, we assume the settings of Chapters 2.2-2.4. In particular, we will frequently refer to the notions of Chapter 2.2.3. Let $\{\psi_j\}_{j\in\mathbb{N}}$ denote an $L^2(\mathbb{R}^3;\mathbb{C})$-orthonormal basis for $H^1(\mathbb{R}^3;\mathbb{C})$. For any $K\in\mathbb{N}$, we define $\mathcal{B}_K:=\{\psi_j\}_{j=1}^K$ and $X_K:=\operatorname{span}\mathcal{B}_K$.

Recall that we denote by $Q\in\mathbb{N}$ the number of electrons in the system under study. Our goal now is to use the sets $\{\mathcal{B}_K\}_{K\in\mathbb{N}}$ to construct a sequence of finite-dimensional, nested subspaces

of the antisymmetric tensor product space $\widehat{\mathscr{H}}^1$ whose union is dense in $\widehat{\mathscr{H}}^1$. To avoid tedious notation in this construction, we will always assume that K is a natural number such that $K \geqslant Q$. Proceeding now, exactly as in Chapter 2.2.3, we first introduce for each such K the index set $\mathcal{I}_K^Q \subset \{1,\ldots,K\}^Q$ given by

$$\mathcal{I}_K^Q := \Big\{ \boldsymbol{\ell} = (\ell_1, \ell_2, \ldots, \ell_Q) \in \{1,\ldots,K\}^Q \colon \ell_1 < \ell_2 < \ldots < \ell_Q \Big\}.$$

Next, we define for each K the set of $\widehat{\mathcal{L}}^2$-orthonormal, Q-particle determinants $\mathcal{B}_K^Q \subset \widehat{\mathscr{H}}^1$ as

$$\mathcal{B}_K^Q := \left\{ \Psi_\alpha(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_Q) = \frac{1}{\sqrt{Q!}} \det\big(\psi_{\alpha_i}(\mathbf{x}_j)\big)_{i,j=1}^Q \colon\ \alpha = (\alpha_1, \alpha_2, \ldots, \alpha_Q) \in \mathcal{I}_K^Q \right\},$$

and we denote, as usual, $\Psi_0(\mathbf{x}_1, \ldots, \mathbf{x}_Q) := \det\big(\psi_i(\mathbf{x}_j)\big)_{i,j=1}^Q$.

It now follows that we can define the sequence $\{\mathcal{V}_K\}_{K \geqslant Q}$ of subspaces of $\widehat{\mathscr{H}}^1$ as $\mathcal{V}_K := \operatorname{span} \mathcal{B}_K^Q$, and it holds that

$$\forall\ K \geqslant Q\colon \quad \dim \mathcal{V}_K = \binom{K}{K-Q}, \qquad \forall\ K_2 > K_1 \geqslant Q\colon \quad \mathcal{V}_{K_1} \subset \mathcal{V}_{K_2} \quad \text{and} \quad \overline{\bigcup_{K \geqslant Q} \mathcal{V}_K}^{\|\cdot\|_{\widehat{\mathscr{H}}^1}} = \widehat{\mathscr{H}}^1.$$

Equipped with the sequence of finite-dimensional subspaces $\{\mathcal{V}_K\}_{K \geqslant Q}$ whose union is dense in $\widehat{\mathscr{H}}^1$, our next task is to introduce a corresponding sequence of finite-dimensional coefficient spaces $\{\mathbb{V}_K\}_{K \geqslant Q}$ whose union is dense in the Hilbert space of sequences $\mathbb{V}$ that was introduced through Definition 2.3.6. To this end, we require some definitions.

Definition 2.5.1 (Excitation Index Sets For Finite Bases)**.**
For each K and each $j \in \{1,\ldots,Q\}$ we define the index set $\mathcal{I}^K$ as

$$\mathcal{I}_j^K := \left\{ \binom{i_1, \ldots, i_j}{\ell_1, \ldots, \ell_j} \colon i_1 < \ldots < i_j \in \{1,\ldots,Q\} \text{ and } \ell_1 < \ldots < \ell_j \in \{Q+1,\ldots,K\} \right\},$$

we set

$$\mathcal{I}^K := \bigcup_{j=1}^Q \mathcal{I}_j^K,$$

and we emphasise that $\bigcup_{K \geqslant Q} \mathcal{I}^K = \mathcal{I}$, i.e., the global excitation index defined through Definition 2.3.1.

Consider Definition 2.5.1 of the excitation index sets $\mathcal{I}_j^K$, $j \in \{1,\ldots,Q\}$. Since each $\mathcal{I}_j^K$ is a subset of the global excitation index set $\mathcal{I}$ defined through Definition 2.3.1, it follows that we can define for any $\mu \in \mathcal{I}_j^K$, excitation and de-excitation operators $\mathcal{X}_\mu \colon \widehat{\mathscr{H}}^1 \to \widehat{\mathscr{H}}^1$ and $\mathcal{X}_\mu^\dagger \colon \widehat{\mathscr{H}}^1 \to \widehat{\mathscr{H}}^1$ through Definitions 2.3.2 and 2.3.3 respectively. Moreover, the results of Theorem 2.3.3 can be applied to these excitation and de-excitation operators, and the following remark summarises some additional properties of these elementary excitation and de-excitation operators.

Remark 2.5.1 (Properties of Excitation and De-excitation Operators Related to the Index Set $\mathcal{I}^K$)**.**
Let the excitation index set $\mathcal{I}^K$ be defined according to Definition 2.5.1. Then the finite-dimensional Q-particle basis $\mathcal{B}_K^Q$ and the finite-dimensional Q-particle approximation space $\mathcal{V}_K$

have the decomposition

$$\mathcal{B}_K^Q := \{\Psi_0\} \cup \{\mathcal{X}_\mu \Psi_0 \colon \ \mu \in \mathcal{I}^K\},$$
$$\mathcal{V}_K := \text{span}\,\{\Psi_0\} \oplus \underbrace{\text{span}\{\mathcal{X}_\mu \Psi_0 \colon \ \mu \in \mathcal{I}^K\}}_{:=\widetilde{\mathcal{V}}_K}.$$

Additionally, for any $\mu, \nu \in \mathcal{I}^K$ and $\kappa \in \mathcal{I} \setminus \mathcal{I}^K$ it holds that

$$\mathcal{X}_\mu \mathcal{X}_\nu \Psi_0 \in \widetilde{\mathcal{V}}_K \qquad \text{and} \qquad \mathcal{X}_\mu^\dagger \mathcal{X}_\nu \Psi_0 \in \widetilde{\mathcal{V}}_K$$
$$\mathcal{X}_\mu \mathcal{X}_\kappa \Psi_0 \notin \mathcal{B}_K^Q \qquad \text{and} \qquad \mathcal{X}_\mu^\dagger \mathcal{X}_\kappa \Psi_0 \notin \mathcal{B}_K^Q,$$
$$\mathcal{X}_\kappa \mathcal{X}_\nu \Psi_0 \notin \mathcal{B}_K^Q \qquad \text{and} \qquad \mathcal{X}_\kappa^\dagger \mathcal{X}_\nu \Psi_0 = 0.$$

Finally, as in Chapter 2.4 we will denote $\Psi_\mu := \mathcal{X}_\mu \Psi_0$ for any $\mu \in \mathcal{I}^K$.

Next we will introduce subspaces of coefficient vectors corresponding to the excitation index sets $\{\mathcal{I}^K\}_{K \geqslant Q}$. The following construction is essentially an adaptation of Definition 2.3.6 of the sequence space $\mathbb{V}$ to finite dimensions.

Definition 2.5.2 (Finite-Dimensional Coefficient Spaces)**.**
Let the excitation index set $\mathcal{I}_K$ be defined through Definition 2.5.1 for $K \geqslant Q$, and let the Hilbert space of sequences $\mathbb{V}$ be defined according to Definition 2.3.6. We define the Hilbert subspace of coefficients $\mathbb{V}_K \subset \mathbb{V}$ as the set

$$\mathbb{V}_K := \left\{\mathbf{t} := (\boldsymbol{t}_\mu)_{\mu \in \mathcal{I}} \in \mathbb{V} \colon \quad \boldsymbol{t}_\mu = 0 \ \forall \mu \notin \mathcal{I}^K\right\}, \tag{2.39}$$

equipped with the $(\cdot,\cdot)_{\mathbb{V}}$ inner product.

Notation 2.5.1. *Consider Definition 2.5.2 of the Hilbert subspace of coefficients $\mathbb{V}_K$, and let $\boldsymbol{t} \in \mathbb{V}_K$ denote an arbitrary element. In the sequel, for clarity of exposition we will frequently denote $\boldsymbol{t} := \boldsymbol{t}_K := \{\boldsymbol{t}_\mu\}_{\mu \in \mathcal{I}^K}$. In other words, by an abuse of notation, we will identify $\mathbb{V}_K$ with the set $\ell^2\left(\mathcal{I}^K\right)$ but equipped with the $(\cdot,\cdot)_{\mathbb{V}}$ inner product.*

As can be expected, the coefficient subspaces $\{\mathbb{V}_K\}_{K \geqslant Q}$ introduced through Definition 2.5.2 inherit many properties from the Q-particle approximation spaces $\{\mathcal{B}_K^Q\}_{K \geqslant Q}$. Indeed, we have the following lemma.

Lemma 2.5.1 (Density of Finite-Dimensional Coefficient Spaces)**.**
Let the infinite-dimensional Hilbert space of sequences $\mathbb{V} \subset \ell^2(\mathcal{I})$ be defined through Definition 2.3.6 and let the Hilbert subspace of coefficients $\mathbb{V}_K \subset \mathbb{V}$ be defined through Definition 2.5.2 for $K \geqslant Q$. Then it holds that

$$\forall\ K_2 > K_1 \geqslant Q \colon \quad \mathbb{V}_{K_1} \subset \mathbb{V}_{K_2} \qquad \text{and} \qquad \overline{\bigcup_{K \geqslant Q} \mathbb{V}_K}^{\|\cdot\|_{\mathbb{V}}} = \mathbb{V}.$$

Proof. The set inclusion is obvious so we focus on proving the density. Note that the density result would also be obvious had the sequence space $\mathbb{V}$ been equipped with the $\|\cdot\|_{\ell^2}$ norm. The density is slightly subtle precisely because we have equipped $\mathbb{V}$ with the non-standard $\|\cdot\|_{\mathbb{V}}$ norm.

Recall that the union of the Q-particle approximation spaces $\{\mathcal{V}_K\}_{K\geqslant Q}$ is dense in $\widehat{\mathcal{H}}^1$. From this we deduce that the union of the Q-particle approximation *subspaces* $\{\widetilde{\mathcal{V}}_K\}_{K\geqslant Q}$ is dense in $\mathrm{span}\{\Psi_0\}^\perp$, where we remind the reader that $\widetilde{\mathcal{V}}_K = \mathrm{span}\{\Psi_\mu\colon\ \mu\in\mathcal{I}^K\} = \{\Psi\in\mathcal{V}_K\colon (\Psi,\Psi_0)_{\widehat{\mathcal{L}}^2} = 0\}$.

Consequently, there exists a sequence of functions $\{\Psi_K\}_{K\geqslant Q}$ with each $\Psi_K := \sum_{\mu\in\mathcal{I}^K} \boldsymbol{t}_\mu^K\Psi_\mu \in \widetilde{\mathcal{V}}_K$ such that $\lim_{K\to\infty}\|\Psi_K - \Psi_{\boldsymbol{s}}^*\|_{\widehat{\mathcal{H}}^1} = 0$. Defining for each $K\geqslant Q$, the sequence $\boldsymbol{t}_K\in\mathbb{V}_K$ as $\boldsymbol{t}_K = \{\boldsymbol{t}_\mu^K\}_{\mu\in\mathcal{I}^K}$, and using the definition of the $\|\cdot\|_{\mathbb{V}}$ norm now yields the required density. □

We are now ready to state the discrete coupled cluster equations corresponding to the approximation spaces we have introduced above. As mentioned at the beginning of this section, these equations are known in the quantum chemical literature as the Full-Coupled Cluster equations in a finite basis.

Full-Coupled Cluster Equations in a Finite Basis:
Let the excitation index set $\mathcal{I}^K$ be defined through Definition 2.5.1 for $K\geqslant Q$, let the Hilbert subspace of coefficients $\mathbb{V}_K\subset\mathbb{V}$ be defined through Definition 2.5.2, and let the coupled cluster function $f\colon\mathbb{V}\to\mathbb{V}^*$ be defined through Definition 2.4.1. We seek a coefficient vector $\boldsymbol{t}_K\in\mathbb{V}_K$ such that for all coefficient vectors $\boldsymbol{s}_K\in\mathbb{V}_K$ it holds that

$$\langle \boldsymbol{s}_K, f(\boldsymbol{t}_K)\rangle_{\mathbb{V}\times\mathbb{V}^*} = 0. \tag{2.40}$$

The remainder of this section will be concerned with the (local) well-posedness analysis of Equation (2.40). We begin with a definition.

Definition 2.5.3 (Restricted Coupled Cluster function on Full-CI spaces)**.**
Let the excitation index set $\mathcal{I}^K$ be defined through Definition 2.5.1 for $K\geqslant Q$ and let the Hilbert subspace of coefficients $\mathbb{V}_K\subset\mathbb{V}$ be defined through Definition 2.5.2. We define the restricted coupled cluster function $f_K\colon\mathbb{V}_K\to\mathbb{V}_K^*$ as the mapping with the property that for all $\boldsymbol{t}_K,\boldsymbol{s}_K\in\mathbb{V}_K$ it holds that

$$\langle \boldsymbol{s}_K, f_K(\boldsymbol{t}_K)\rangle_{\mathbb{V}_K\times\mathbb{V}_K^*} := \langle \boldsymbol{s}_K, f(\boldsymbol{t}_K)\rangle_{\mathbb{V}\times\mathbb{V}^*}\,.$$

It is readily seen that solutions $\boldsymbol{t}_K^*\in\mathbb{V}_K$ to the Full-CC equations in a finite basis (2.40) are nothing else than zeros of the restricted coupled cluster function $f_K\colon\mathbb{V}_K\to\mathbb{V}_K^*$ defined through Definition 2.5.3. The following result, whose proof can, for instance, be found in [87], is essentially a finite-dimensional analogue of Theorem 2.4.1 and establishes a relationship between these zeros of the restricted coupled cluster function and intermediately normalised eigenfunctions of the Full-CI Hamiltonian $H_K\colon\mathcal{V}_K\to\mathcal{V}_K^*$ defined through Equation (2.7).

Theorem 2.5.2 (Relation between Restricted Coupled Cluster Zeros and Full-CI Eigenfunctions)**.**
Let the restricted coupled cluster function $f_K\colon\mathbb{V}_K\to\mathbb{V}_K^*$ be defined through Definition 2.5.3, and let the Full-CI Hamiltonian $H_K\colon\mathcal{V}_K\to\mathcal{V}_K^*$ be defined through Equation (2.7)*. Then*

1. *For any zero $\boldsymbol{t}_K^* = \{\boldsymbol{t}_\mu^*\}_{\mu\in\mathcal{I}^K}\in\mathbb{V}_K$ of the restricted CC function, the function $\Psi_K^* = e^{\mathcal{T}_K^*}\Psi_0\in\mathcal{V}_K$ with $\mathcal{T}_K^* = \sum_{\mu\in\mathcal{I}^K}\boldsymbol{t}_\mu^*\mathcal{X}_\mu$ is an intermediately normalised eigenfunction of the Full-CI Hamiltonian. Moreover, the eigenvalue corresponding to the eigenfunction Ψ_K^* coincides with the discrete CC energy $\mathcal{E}_{K,\mathrm{CC}}^*$ generated by $\boldsymbol{t}_K^*$ as defined through Equation* (2.16)*.*

2. *Conversely, for any intermediately normalised eigenfunction $\Psi_K^* \in \mathcal{V}_K$ of the Full-CI Hamiltonian, there exists $\boldsymbol{t}_K^* = \{\boldsymbol{t}_\mu^*\}_{\mu\in\mathcal{G}^K} \in \mathbb{V}_K$ such that $\boldsymbol{t}_K^*$ is a zero of the restricted CC function and $\Psi_K^* = e^{\mathcal{T}_K^*}\Psi_0 \in \mathcal{V}_K$ with $\mathcal{T}_K^* = \sum_{\mu\in\mathcal{G}^K} \boldsymbol{t}_\mu^* \mathcal{X}_\mu$. Moreover, the discrete CC energy $\mathcal{E}_{K,\mathrm{CC}}^*$ generated by $\boldsymbol{t}_K^*$ through through Equation* (2.16) *coincides with the eigenvalue corresponding to the eigenfunction Ψ_K^*.*

In view of Theorem 2.5.2, the goal of our analysis in this section will be two-fold: first, we would like to demonstrate, exactly as in the infinite-dimensional case, that solutions $\boldsymbol{t}_K^* \in \mathbb{V}_K$ of the Full-CC equations (2.40) that correspond to non-degenerate eigenpairs of the Full-CI Hamiltonian are locally unique. Second, we wish to obtain a characterisation of the error between solutions $\boldsymbol{t}_K^* \in \mathbb{V}_K$ of the Full-CC equations (2.40) and solutions $\boldsymbol{t}^* \in \mathbb{V}$ of the continuous coupled cluster equations (2.17). For the latter analysis we will appeal to classical results from the numerical analysis of Galerkin discretisations of non-linear equations but the former task is essentially trivial since the Full-CC equations have the same structure as the continuous CC equations and hence our proofs from Chapter 2.4 can be copied with minor amendments. For the sake of brevity therefore, we simply state the final result on local uniqueness of solutions to the Full-CC equations in a finite basis (2.40).

Theorem 2.5.3 (Local Well-Posedness of the Full-Coupled Cluster Equations in a Finite Basis)**.**

Let $\mathbb{V}_K \subset \mathbb{V}$ denote the Hilbert subspace of coefficients as defined through Definition 2.5.2 for $K \geqslant Q$, let the restricted coupled cluster function $f_K\colon \mathbb{V}_K \to \mathbb{V}_K^*$ be defined through Definition 2.5.3, let $\boldsymbol{t}_K^* := \{\boldsymbol{t}_\mu^*\}_{\mu\in\mathcal{G}^K} \in \mathbb{V}_K$ denote a zero of the restricted coupled cluster function corresponding to any intermediately normalised eigenfunction $\Psi_K^* \in \mathcal{V}_K$ of the Full-CI Hamiltonian $H_K\colon \mathcal{V}_K \to \mathcal{V}_K^*$ with non-degenerate eigenvalue $\mathcal{E}_K^*$, let $\gamma_K > 0$ denote the inf-sup constant of the shifted Full-CI Hamiltonian $H_K - \mathcal{E}_K^*$ on $\{\Psi_K^*\}^\perp \subset \mathcal{V}_K$, let $\Theta_K > 0$ be defined as $\Theta_K := \|e^{(\mathcal{T}_K^*)^\dagger}\|_{\mathcal{V}_K\to\mathcal{V}_K}\|\mathbb{P}_0^\perp e^{-\mathcal{T}_K^*}\|_{\mathcal{V}_K\to\mathcal{V}_K}$ with $\mathcal{T}_K^* := \sum_{\mu\in\mathcal{G}^K} \boldsymbol{t}_\mu^*\mathcal{X}_\mu$, let the continuity constant $\alpha_{\boldsymbol{t}_K^*} > 0$ and the Lipschitz continuity function $L_{\boldsymbol{t}_K^*}\colon \mathbb{R}_+ \to \mathbb{R}_+$ be defined according to Notation 2.4.2, and define the constant

$$\mathrm{T} := \mathrm{T}(K) := \min_{\delta>0}\left\{\delta,\ \frac{\gamma_K}{L_{\boldsymbol{t}_K^*}(\delta)\Theta_K},\ 2\frac{\alpha_{\boldsymbol{t}_K^*}}{L_{\boldsymbol{t}_K^*}(\delta)}\right\}.$$

Then $f_K\big(\mathrm{B}_\mathrm{T}(\boldsymbol{t}_K^*)\big)$ is an open subset of $\mathbb{V}_K^*$, the restriction of f_K to $\mathrm{B}_\mathrm{T}(\boldsymbol{t}_K^*)$ is a diffeomorphism*, and for all $\boldsymbol{s}_K \in \mathrm{B}_\mathrm{T}(\boldsymbol{t}_K^*)$ we have the error estimate*

$$\frac{1}{2}\frac{1}{\alpha_{\boldsymbol{t}_K^*}}\|f_K(\boldsymbol{s}_K)\|_{\mathbb{V}_K^*} \leqslant \|\boldsymbol{t}_K^* - \boldsymbol{s}_K\|_{\mathbb{V}_K} \leqslant 2\frac{\Theta_K}{\gamma_K}\|f_K(\boldsymbol{s}_K)\|_{\mathbb{V}_K^*}. \tag{2.41}$$

In particular, $\boldsymbol{t}_K^*$ is the unique solution of the Full-Coupled Cluster equations in a finite basis (2.13) *in the open ball $\mathrm{B}_{\mathrm{T}_K}(\boldsymbol{t}_K^*)$.*

Proof. The proof is essentially identical to the proof of Theorem 2.4.4 with some obvious modifications. We first obtain an expression for the Full-CC Jacobian $\mathrm{D}f_K(\boldsymbol{t}_K)\colon \mathbb{V}_K \to \mathbb{V}_K^*$ at any $\boldsymbol{t}_K \in \mathbb{V}_K$ exactly as in the infinite-dimensional case. Thanks to Theorem 2.5.2, we can deduce from this expression that the Jacobian $\mathrm{D}f_K(\boldsymbol{t}_K^*)$ at any zero $\boldsymbol{t}_K^* \in \mathbb{V}_K$ of the restricted CC function has the form

$$\langle \boldsymbol{w}_K, \mathrm{D}f_K(\boldsymbol{t}_K^*)\boldsymbol{s}_K\rangle_{\mathbb{V}_K\times\mathbb{V}_K^*} = \left\langle \sum_{\mu\in\mathcal{G}^K} \boldsymbol{w}_\mu\mathcal{X}_\mu\Psi_0, e^{-\mathcal{T}_K^*}(H-\mathcal{E}_K^*)e^{\mathcal{T}_K^*}\sum_{\nu\in\mathcal{G}^K}\boldsymbol{s}_\nu\mathcal{X}_\nu\Psi_0\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}},$$

for all $\boldsymbol{s}_K, \boldsymbol{w}_K \in \mathbb{V}_K$ with $\boldsymbol{s}_K = \{\boldsymbol{s}_\nu\}_{\nu\in\mathcal{G}^K}$ and $\boldsymbol{w}_K = \{\boldsymbol{w}_\mu\}_{\mu\in\mathcal{G}^K}$ where $\mathcal{T}_K^* := \sum_{\mu\in\mathcal{G}^K} \boldsymbol{t}_\mu^* \mathcal{X}_\mu$.

In analogy with Definition 2.4.2, we can then introduce an operator $\mathcal{A}_K(\boldsymbol{t}_K^*)\colon \widetilde{\mathcal{V}}_K \to \widehat{\mathcal{H}}^{-1}$ that characterises the action of the Full-CC derivative $\mathrm{D}f_K(\boldsymbol{t}_K^*)$ and show that this operator is an isomorphism from $\widetilde{\mathcal{V}}_K$ to $\widetilde{\mathcal{V}}_K^*$ exactly as in Theorem 2.4.3. The local-uniqueness result then readily follows. □

Consider the setting of Theorem 2.5.3. For very small molecules discretised in minimal basis sets, it is possible to perform Full-CI calculations and thereby gain access to the derivative $\mathrm{D}f_K(\boldsymbol{t})$ of the restricted coupled cluster function at $\boldsymbol{t} = \boldsymbol{t}_{\mathrm{FCI}}^* \in \mathbb{V}_K$, i.e., at the coefficient vector $\boldsymbol{t}_{\mathrm{FCI}}^*$ which generates the Full-CI ground state wave-function. It is natural to ask how the bounds that we have derived compare to the exact norm of the inverse $\mathrm{D}f_K^{-1}(t_{\mathrm{FCI}}^*)$. While a comprehensive numerical study involving state-of-the-art quantum chemistry basis sets for moderately large molecules, is computationally unfeasible, there is some hope that numerical experiments can be performed on certain very small molecules using so-called minimal basis sets. A numerical study of this nature is left to a future contribution but some preliminary numerical results are given in Table 2.2 and Figures 2.1 and 2.2. Based on these results, the lower bounds that we have derived for the operator norm $\|\mathrm{D}f_K^{-1}(\boldsymbol{t}_{\mathrm{FCI}}^*)\|^{-1}_{\mathbb{V}_K^*\to\mathbb{V}_K}$ seem reasonable at equilibrium but tend to degrade in the bond dissociation regime.

Molecule	$\|\boldsymbol{t}_{\mathrm{FCI}}^*\|_{\mathbb{V}_K}$	Monotonicity constant Γ from Eq. (2.19a)	$\|\mathrm{D}f_K^{-1}(t_{\mathrm{FCI}}^*)\|^{-1}_{\mathbb{V}_K^*\to\mathbb{V}_K}$	γ_K/Θ_K	$\dfrac{\gamma_K/\Theta_K}{\|\mathrm{D}f_K^{-1}(t_{\mathrm{FCI}}^*)\|^{-1}_{\mathbb{V}_K^*\to\mathbb{V}_K}}$
BeH2	0.2343	0.0363	0.3379	0.2568	0.7599
BH3	0.2844	-0.0950	0.3060	0.2081	0.6801
HF	0.2038	-0.0083	0.2995	0.2529	0.8444
H2O	0.2687	0.0249	0.3576	0.2789	0.7799
LiH	0.1792	-0.0065	0.2628	0.2164	0.8234
NH3	0.3074	-0.0325	0.4113	0.2784	0.6769

Table 2.2: Examples of numerically computed constants for a collection of small molecules at equilibrium geometries. The calculations were performed in STO-6G basis sets with the exception of the HF and LiH molecules for which 6-31G basis sets were used. To simplify calculations, the canonical $\widehat{\mathcal{H}}^1$ norm was replaced with an equivalent norm induced by the mean-field Hartree Fock operator (see, e.g., [87]).

We now turn to the second goal of this section, namely to a study of the error between solutions $\boldsymbol{t}_K^* \in \mathbb{V}_K$ of the Full-CC equations (2.40) and solutions $\boldsymbol{t}^* \in \mathbb{V}$ of the continuous coupled cluster equations (2.17). Since the Full-CC equations are simply Galerkin discretisations of the continuous CC equations, their local well-posedness can be deduced from classical results in non-linear numerical analysis. Indeed, we merely have to obtain an appropriate invertibility result for the coupled cluster Fréchet derivative restricted to the coefficient subspaces $\{\mathbb{V}_K\}_{K\geqslant Q}$ and we must establish that the subspaces $\{\mathbb{V}_K\}_{K\geqslant Q}$ have the approximation property with respect to $\mathbb{V}$. The latter demonstration is a simple consequence of the density of $\bigcup_{K\geqslant Q} \mathbb{V}_K$ in $\mathbb{V}$ which has already been proven in Lemma 2.5.1. We therefore focus on obtaining the required invertibility result.

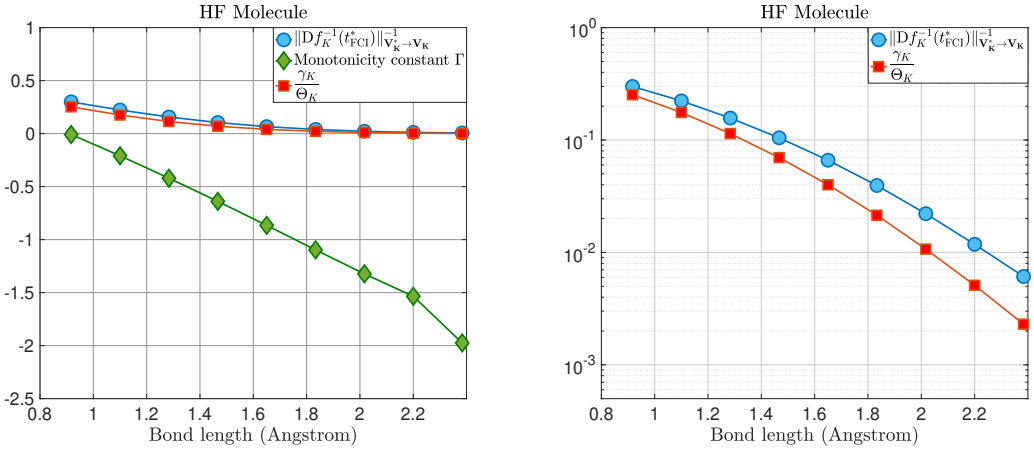


Figure 2.1: Numerically computed constants for the HF molecule at different bond lengths. The equilibrium bond length is 0.9168 Angstrom. The figure on the right uses a log scale on the y-axis.

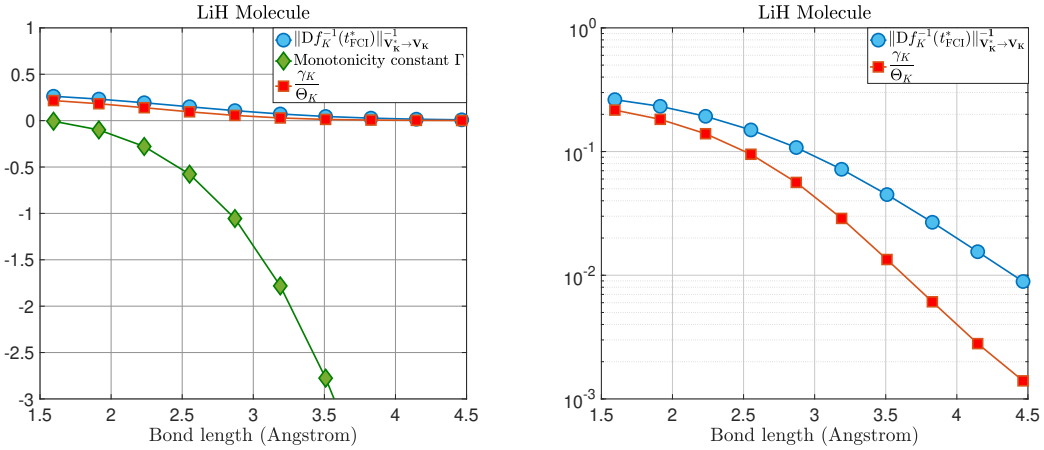


Figure 2.2: Numerically computed constants for the LiH molecule at different bond lengths. The equilibrium bond length is 1.5949 Angstrom. The figure on the right uses a log scale on the y-axis.

We begin by defining projection operators corresponding to the various finite-dimensional approximation spaces we have introduced.

Definition 2.5.4 (Projection Operators)**.**

Let $\mathcal{V}_K = \operatorname{span}\{\Psi_0\} \cup \widetilde{\mathcal{V}}_K \subset \widehat{\mathcal{H}}^1$ denote the finite-dimensional Q-particle approximation space for $K \geqslant Q$ and let $\mathbb{V}_K \subset \mathbb{V}$ denote the Hilbert subspace of coefficients as defined through Definition 2.5.2. Then

- *We denote by $\mathbb{P}_K \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^1$, the $\widehat{\mathcal{H}}^1$-orthogonal projection operator onto $\mathcal{V}_K$ and by $\mathbb{P}_K^{\perp}$*

its complement, i.e., $\mathbb{P}_K^\perp = \mathbb{I} - \mathbb{P}_K$.

- *We denote by* $\Pi_K \colon \mathbb{V} \to \mathbb{V}$, *the* $(\cdot,\cdot)_{\mathbb{V}}$*-orthogonal projection operator onto* $\mathbb{V}_K$ *and by* $\Pi_K^\perp$ *its complement, i.e.,* $\Pi_K^\perp = \mathbb{I} - \Pi_K$.

Notation 2.5.4 (Cluster Operators Involving Projections)**.**
Consider the setting of Definition 2.5.4 and let $\boldsymbol{t} \in \mathbb{V}$. *In the sequel, we will frequently consider cluster operators generated by* $\Pi_K \boldsymbol{t}$ *or* $\Pi_K^\perp \boldsymbol{t}$. *We will therefore use the notation* $\mathcal{T}(\Pi_K)$ *and* $\mathcal{T}(\Pi_K^\perp)$ *respectively to denote these cluster operators, i.e., we denote*

$$\mathcal{T}(\Pi_K) := \sum_{\mu \in \mathcal{G}^K} \boldsymbol{s}_\mu \mathcal{X}_\mu \quad \text{where} \quad \{\boldsymbol{s}_\mu\}_{\mu \in \mathcal{G}^K} = \Pi_K \boldsymbol{t} \in \mathbb{V}_K, \quad \text{and}$$
$$\mathcal{T}(\Pi_K^\perp) := \sum_{\mu \in \mathcal{G}} \boldsymbol{r}_\mu \mathcal{X}_\mu \quad \text{where} \quad \{\boldsymbol{r}_\mu\}_{\mu \in \mathcal{G}} = \Pi_K^\perp \boldsymbol{t} \in \mathbb{V}.$$

We are now ready to state the main technical lemma of this section. We emphasise that the proof of this lemma assumes that any isolated, simple eigenpair of the electronic Hamiltonian can be approximated by a sequence of simple eigenpairs of the Full-CI Hamiltonian.

Lemma 2.5.2 (Invertibility of the coupled cluster Fréchet derivative on $\mathbb{V}_K$)**.**
Let the coupled cluster function $f \colon \mathbb{V} \to \mathbb{V}^*$ *be defined through Definition 2.4.1, for any* $\boldsymbol{t} \in \mathbb{V}$ *let* $\mathrm{D}f(\boldsymbol{t})$ *denote the Fréchet derivative of the coupled cluster function as defined through Equation* (2.20)*, let* $\boldsymbol{t}^* \in \mathbb{V}$ *denote a zero of the coupled cluster function corresponding to an intermediately normalised eigenfunction* $\Psi^* \in \widehat{\mathcal{H}}^1$ *of the electronic Hamiltonian* $H \colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ *with isolated, non-degenerate eigenvalue* $\mathcal{E}^*$, *let* $\mathbb{V}_K \subset \mathbb{V}$ *denote the Hilbert subspace of coefficients as defined through Definition 2.5.2 for* $K \geqslant Q$, *let the Full-CI Hamiltonian* $H_K \colon \mathcal{V}_K \to \mathcal{V}_K^*$ *be defined according to Definition 2.7 and let* $(\Psi_K^*, \mathcal{E}_K^*) \in \mathcal{V}_K \times \mathbb{R}$ *be a sequence of simple eigenpairs of the Full-CI Hamiltonians* $\{H_K\}_{K \geqslant Q}$, *i.e.,*

$$\forall \Phi_K \in \mathcal{V}_K \colon \qquad \langle \Phi_K, H_K \Psi_K^* \rangle_{\mathcal{V}_K \times \mathcal{V}_K^*} = \mathcal{E}_K^* \langle \Phi_K, \Psi_K^* \rangle_{\widehat{\mathcal{H}}^1 \times \widehat{\mathcal{H}}^{-1}} \quad \text{with} \quad \mathcal{E}_K^* \text{ simple} \quad \text{and such that} \tag{2.42}$$

$$\lim_{K \to \infty} \|\Psi^* - \Psi_K^*\|_{\widehat{\mathcal{H}}^1} = 0, \quad \lim_{K \to \infty} |\mathcal{E}^* - \mathcal{E}_K^*| = 0.$$

Then for all K *sufficiently large, there exist a constant* $\gamma_K > 0$ *uniformly bounded below in* K, *a constant* $\Theta_K > 0$ *uniformly bounded above in* K, *a constant* $\varepsilon_K > 0$ *such that* $\lim_{K \to \infty} \varepsilon_K = 0$, *a constant* $\omega_K > 0$ *such that* $\lim_{K \to \infty} \omega_K = 1$, *and we have the estimate*

$$\inf_{0 \neq \boldsymbol{w}_K \in \mathbb{V}_K} \sup_{0 \neq \boldsymbol{s}_K \in \mathbb{V}_K} \frac{\langle \boldsymbol{w}_K, \mathrm{D}f(\boldsymbol{t}^*) \boldsymbol{s}_K \rangle_{\mathbb{V} \times \mathbb{V}^*}}{\|\boldsymbol{w}_K\|_{\mathbb{V}} \|\boldsymbol{s}_K\|_{\mathbb{V}}} \geqslant \frac{\gamma_K / \omega_K - \varepsilon_K}{\Theta_K}.$$

Proof. Firstly, let us remark that the existence of convergent sequence of simple eigenpaires $(\Psi_K^*, \mathcal{E}_K^*) \in \mathcal{V}_K \times \mathbb{R}$ satisfying Equation (2.42) is guaranteed by well-known approximability results for linear eigenvalue problems (see, e.g., [20, Example 5.9]).

Let $\boldsymbol{w}_K = \{\boldsymbol{w}_\mu\}_{\mu\in\mathcal{G}^K}, \boldsymbol{s}_K = \{\boldsymbol{s}_\mu\}_{\mu\in\mathcal{G}^K} \in \mathbb{V}_K$ be arbitrary and let the bounded linear operator $\mathcal{A}(\boldsymbol{t}^*)\colon \widetilde{\mathcal{V}} \to \widetilde{\mathcal{V}}^*$ be defined according to Definition 2.4.2. It follows from Corollary 2.4.1.1 that

$$\langle \boldsymbol{w}_K, \mathrm{D}f(\boldsymbol{t}^*)\boldsymbol{s}_K\rangle_{\mathbb{V}\times\mathbb{V}^*} = \langle \mathcal{W}_K\Psi_0, \mathcal{A}(\boldsymbol{t}^*)\mathcal{S}_K\Psi_0\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} = \left\langle \mathcal{W}_K\Psi_0, e^{-\mathcal{T}^*}\left(H-\mathcal{E}^*\right)e^{\mathcal{T}^*}\mathcal{S}_K\Psi_0\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}},$$

where $\mathcal{W}_K := \sum_{\mu\in\mathcal{G}^K}\boldsymbol{w}_\mu\mathcal{X}_\mu$ and $\mathcal{S}_K := \sum_{\mu\in\mathcal{G}^K}\boldsymbol{s}_\mu\mathcal{X}_\mu$. To avoid tedious notation, let us define $\Phi_{\mathcal{W}} := \mathcal{W}_K\Psi_0 \in \widetilde{\mathcal{V}}_K$ and $\Phi_{\mathcal{S}} := \mathcal{S}_K\Psi_0 \in \widetilde{\mathcal{V}}_K$. Obviously, we now have

$$\left\langle \Phi_{\mathcal{W}}, e^{-\mathcal{T}^*}\left(H-\mathcal{E}^*\right)e^{\mathcal{T}^*}\Phi_{\mathcal{S}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} = \left\langle e^{\mathcal{T}^*}\Phi_{\mathcal{S}}, \left(H-\mathcal{E}^*\right)e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}. \tag{2.43}$$

Since Ψ^* is intermediately normalisable by assumption, there exists $\widetilde{K_0} \in \mathbb{N}$ such that for all $K \geqslant \widetilde{K_0}$, the eigenfunction $\Psi_K^* \in \mathcal{V}_K$ is intermediately normalisable. In the remainder of this proof, we assume that indeed $K \geqslant \widetilde{K_0}$ and we denote by $\boldsymbol{t}_K^* := \{\boldsymbol{t}_\mu^*\}_{\mu\in\mathcal{G}} \in \mathbb{V}_K$ the coefficient vector with the property that $\Psi_K^* = e^{\mathcal{T}_K^*}\Psi_0$ where $\mathcal{T}_K^* := \sum_{\mu\in\mathcal{G}^K}\boldsymbol{t}_\mu^*\mathcal{X}_\mu$. Let us emphasise here that since Ψ_K^* is an eigenfunction of the Full-CI Hamiltonian, it follows from Theorem 2.5.2 that $\boldsymbol{t}_K^*$ is a zero of the restricted coupled cluster function $f_K\colon \mathbb{V}_K \to \mathbb{V}_K^*$ defined through Definition 2.5.3.

Recalling now that $\Phi_{\mathcal{S}}$ is arbitrary (due to the fact that the sequence $\boldsymbol{s}_K = \{\boldsymbol{s}_\mu\}_{\mu\in\mathcal{G}^K} \in \mathbb{V}_K$ was chosen arbitrarily), we may in particular set for any $\Phi_{K,\perp}^* \in \{\Psi_K^*\}^\perp \subset \mathcal{V}_K$:

$$\Phi_{\mathcal{S}} := \mathbb{P}_0^\perp e^{-\mathcal{T}^*(\Pi_K)}\Phi_{K,\perp}^* \in \widetilde{\mathcal{V}}_K,$$

where $\mathcal{T}^*(\Pi_K)$ denotes the cluster operator generated by $\Pi_K\boldsymbol{t}^* \in \mathbb{V}_K$ (recall Notation 2.5.4) and we have used the fact that, thanks to the properties of the excitation operators $\{\mathcal{X}_\mu\}_{\mu\in\mathcal{G}}$ given by Remark 2.5.1, it holds that $e^{-\mathcal{T}^*(\Pi_K)}\Psi_K \in \mathcal{V}_K$ for any $\Psi_K \in \mathcal{V}_K$.

Plugging in this choice of $\Phi_{\mathcal{S}}$ in Equation (2.43) now yields

$$\begin{aligned}\langle \mathcal{W}_K\Psi_0, \mathcal{A}(\boldsymbol{t}^*)\mathcal{S}_K\Psi_0\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} = &\underbrace{\left\langle e^{\mathcal{T}^*}e^{-\mathcal{T}^*(\Pi_K)}\Phi_{K,\perp}^*, \left(H-\mathcal{E}^*\right)e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{I})}\\ &-\underbrace{\left\langle e^{\mathcal{T}^*}\mathbb{P}_0e^{-\mathcal{T}^*(\Pi_K)}\Phi_{K,\perp}^*, \left(H-\mathcal{E}^*\right)e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{II})}.\end{aligned} \tag{2.44}$$

We claim that the term (II) is identically zero. Indeed, using the fact that $e^{\mathcal{T}^*}\Psi_0 = \Psi^*$ by assumption, a straightforward calculation shows that

$$(\mathrm{II}) = \left(\Psi_0, e^{-\mathcal{T}^*(\Pi_K)}\Phi_{K,\perp}^*\right)_{\widehat{\mathcal{L}}^2}\left\langle e^{\mathcal{T}^*}\Psi_0, \left(H-\mathcal{E}^*\right)e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}},$$

and the second term in the product above is zero as $He^{\mathcal{T}^*}\Psi_0 = \mathcal{H}\Psi^* = \mathcal{E}^*\Psi^*$.

It therefore remains to simplify the term (I). To this end, we observe that we can write

$$\begin{aligned}(\mathrm{I}) &= \left\langle e^{\mathcal{T}^*(\Pi_K^\perp)}\Phi^*_{K,\perp}, (H-\mathcal{E}^*)\, e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}} \\ &= \underbrace{\left\langle \Phi^*_{K,\perp}, (H-\mathcal{E}^*)\, e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{IA})} \\ &\quad + \underbrace{\left\langle \left(e^{\mathcal{T}^*(\Pi_K^\perp)}-\mathbb{I}\right)\Phi^*_{K,\perp}, (H-\mathcal{E}^*)\, e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{IB})}.\end{aligned}$$

Focusing first on the term (IB) and using the Cauchy-Schwarz inequality, we may write

$$(\mathrm{IB}) \geqslant -\left\|e^{\mathcal{T}^*(\Pi_K^\perp)}-\mathbb{I}\right\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}\|H-\mathcal{E}^*\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^{-1}}\left\|\Phi^*_{K,\perp}\right\|_{\widehat{\mathcal{H}}^1}\left\|e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1}. \tag{2.45}$$

We now claim that in fact

$$\lim_{K\to\infty}\left\|e^{\mathcal{T}^*(\Pi_K^\perp)}-\mathbb{I}\right\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}=0.$$

Indeed, thanks to the boundedness properties of the excitation operators given in Theorem 2.3.3, it holds that

$$\|e^{-\mathcal{T}^*(\Pi_K)}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}\leqslant e^{\|\mathcal{T}^*(\Pi_K)\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}}\leqslant e^{\beta\|\Pi_K\boldsymbol{t}^*\|_{\mathbb{V}}}\leqslant e^{\beta\|\boldsymbol{t}^*\|_{\mathbb{V}}},$$

where the constant $\beta>0$ depends only on Q.

Therefore, we need only show that $\lim_{K\to\infty}\|e^{\mathcal{T}^*}-e^{\mathcal{T}^*(\Pi_K)}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}=0$. Recall however, that the exponential function is of class $\mathscr{C}^\infty$ on the algebra of bounded operators on $\widehat{\mathcal{H}}^1$, and thus it suffices to show that

$$\lim_{K\to\infty}\|\mathcal{T}^*-\mathcal{T}^*(\Pi_K)\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}=0.$$

But this is an obvious consequence of the density of the coefficient spaces $\{\mathbb{V}_K\}_{K\geqslant Q}$ in $\mathbb{V}$. Indeed,

$$\lim_{K\to\infty}\|\mathcal{T}^*-\mathcal{T}^*(\Pi_K)\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}\leqslant\beta\lim_{K\to\infty}\|\boldsymbol{t}^*-\Pi_K\boldsymbol{t}^*\|_{\mathbb{V}}=0. \tag{2.46}$$

Consequently, combining Equations (2.45) and (2.46), we obtain the existence of a constant $\varepsilon_{1,K}>0$ with the property that $\lim_{K\to\infty}\varepsilon_{1,K}=0$ and such that

$$(\mathrm{IB})\geqslant-\varepsilon_{1,K}\left\|\Phi^*_{K,\perp}\right\|_{\widehat{\mathcal{H}}^1}\left\|e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1}. \tag{2.47}$$

Let us now return to the term (IA). Notice that we may write

$$(\mathrm{IA})=\underbrace{\left\langle\Phi^*_{K,\perp}, (H-\mathcal{E}^*)\, e^{-(\mathcal{T}_K^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{IAA})}+\underbrace{\left\langle\Phi^*_{K,\perp}, (H-\mathcal{E}^*)\left(e^{-(\mathcal{T}^*)^\dagger}-e^{-(\mathcal{T}_K^*)^\dagger}\right)\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}_{:=(\mathrm{IAB})},$$

where we recall from Equation (2.42) that $\boldsymbol{t}_K^*=\{\boldsymbol{t}_\mu^*\}_{\mu\in\mathcal{I}_K}\in\mathbb{V}_K$ is the coefficient vector such that $e^{\mathcal{T}_K^*}\Psi_0=\Psi_K^*\in\mathcal{V}_K$.

We first simplify the term (IAB). Thanks to the Cauchy-Schwarz inequality we may write

$$(\mathrm{IAB}) \geqslant -\left\|e^{-(\mathcal{T}^*)^\dagger} - e^{-(\mathcal{T}_K^*)^\dagger}\right\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1} \left\|e^{(\mathcal{T}^*)^\dagger}\right\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1} \|H-\mathcal{E}^*\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^{-1}} \left\|\Phi_{K,\perp}^*\right\|_{\widehat{\mathcal{H}}^1} \left\|e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1}. \tag{2.48}$$

We now claim that in fact $\lim_{K\to\infty}\left\|e^{-(\mathcal{T}^*)^\dagger} - e^{-(\mathcal{T}_K^*)^\dagger}\right\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1} = 0$. Indeed, an easy calculation using the continuity properties of cluster operators given by Theorem 2.3.3, shows the existence of a constant $\widetilde{\beta^\dagger} > 0$, depending only on Q, such that for any $K \geqslant Q$ it holds that

$$\left\|e^{-(\mathcal{T}^*)^\dagger} - e^{-(\mathcal{T}_K^*)^\dagger}\right\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1} \leqslant \widetilde{\beta^\dagger}\left\|e^{\mathcal{T}^*}\Psi_0 - e^{\mathcal{T}_K^*}\Psi_0\right\|_{\widehat{\mathcal{H}}^1} = \|\Psi^* - \Psi_K^*\|_{\widehat{\mathcal{H}}^1}. \tag{2.49}$$

The claim now follows by using the convergence of the approximate eigenvector $\Psi_K^* \in \mathcal{V}_K$ to $\Psi^* \in \widehat{\mathcal{H}}^1$ from Equation (2.42). Consequently, we obtain the existence of a constant $\varepsilon_{2,K} > 0$ with the property that $\lim_{K\to\infty}\varepsilon_{2,K} = 0$ and such that

$$(\mathrm{IB}) \geqslant -\varepsilon_{2,K}\left\|\Phi_{K,\perp}^*\right\|_{\widehat{\mathcal{H}}^1}\left\|e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1}. \tag{2.50}$$

Focusing finally on the term (IAA), a simple calculation shows that for any $\Phi_{\mathcal{W}} \in \widetilde{\mathcal{V}}_K$, we have that $e^{-(\mathcal{T}_K^*)^\dagger}\Phi_{\mathcal{W}} \in \{\Psi_K^*\}^\perp \subset \mathcal{V}_K$. Furthermore, $\mathcal{E}^*$ is a simple, isolated eigenvalue by assumption and $\lim_{K\to\infty}\mathcal{E}_K^* = \mathcal{E}^*$. Since $\Phi_{K,\perp}^* \in \{\Psi_K^*\}^\perp \subset \mathcal{V}_K$ is arbitrary, we therefore deduce the existence of $\widehat{K_0} \in \mathbb{N}$ sufficiently large such that for all $K \geqslant \widehat{K_0}$ the shifted Full-CI Hamiltonian $H_K - \mathcal{E}^*$ satisfies an inf-sup condition on $\{\Psi_K^*\}^\perp \subset \mathcal{V}_K$, and as a consequence,

$$\sup_{0\neq\Phi_{K,\perp}^*\in\{\Psi_K^*\}^\perp} \frac{(\mathrm{IAA})}{\|\Phi_{K,\perp}^*\|_{\widehat{\mathcal{H}}^1}} = \sup_{0\neq\Phi_{K,\perp}^*\in\{\Psi_K^*\}^\perp} \frac{\left\langle\Phi_{K,\perp}^*, (H-\mathcal{E}^*)\, e^{-(\mathcal{T}_K^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}{\|\Phi_{K,\perp}^*\|_{\widehat{\mathcal{H}}^1}} \geqslant \gamma_K\left\|e^{-(\mathcal{T}_K^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1}, \tag{2.51}$$

where γ_K denotes the inf-sup constant of the shifted Full-CI Hamiltonian $H_K - \mathcal{E}^*$ on $\{\Psi_K^*\}^\perp \subset \mathcal{V}_K$ for $K \geqslant \widehat{K_0}$. For the remainder of this proof, we assume that indeed $K \geqslant \widehat{K_0}$.

Notice that this last bound can be written as

$$\gamma_K\left\|e^{-(\mathcal{T}_K^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1} = \gamma_K\left\|e^{-(\mathcal{T}_K^*-\mathcal{T}^*)^\dagger}e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1} \geqslant \frac{\gamma_K}{\|e^{(\mathcal{T}_K^*-\mathcal{T}^*)^\dagger}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}}\left\|e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1},$$

where the inequality follows from the invertibility of the exponential map. Using now a similar calculation to the one used to obtain Inequality (2.49), it can easily be shown that $\lim_{K\to\infty}\|e^{(\mathcal{T}_K^*-\mathcal{T}^*)^\dagger}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1} = 1$. Consequently, we obtain the existence of constant $\omega_K > 0$ with the property that $\lim_{K\to\infty}\omega_K = 1$ and such that

$$\sup_{0\neq\Phi_{K,\perp}^*\in\{\Psi_K^*\}^\perp} \frac{(\mathrm{IAA})}{\|\Phi_{K,\perp}^*\|_{\widehat{\mathcal{H}}^1}} = \sup_{0\neq\Phi_{K,\perp}^*\in\{\Psi_K^*\}^\perp} \frac{\left\langle\Phi_{K,\perp}^*, (H-\mathcal{E}^*)\, e^{-(\mathcal{T}_K^*)^\dagger}\Phi_{\mathcal{W}}\right\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}{\|\Phi_{K,\perp}^*\|_{\widehat{\mathcal{H}}^1}} \geqslant \frac{\gamma_k}{\omega_k}\left\|e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\right\|_{\widehat{\mathcal{H}}^1}. \tag{2.52}$$

Combining now the estimates (2.43)-(2.52) allows us to conclude that

$$\begin{aligned}\sup_{0\neq \boldsymbol{s}_K\in\mathbb{V}_K}\frac{\langle \boldsymbol{w}_K, \mathrm{D}f(\boldsymbol{t}^*)\boldsymbol{s}_K\rangle_{\mathbb{V}\times\mathbb{V}^*}}{\|\boldsymbol{s}_K\|_{\mathbb{V}}} &= \sup_{0\neq\Phi_{\mathcal{S}}\in\widetilde{\mathcal{V}}_K}\frac{\langle \mathcal{W}_K\Psi_0, \mathcal{A}(\boldsymbol{t}^*)\mathcal{S}_K\Psi_0\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}}{\|\Phi_{\mathcal{S}}\|_{\widehat{\mathcal{H}}^1}}\\ &\geqslant \sup_{0\neq\Phi^*_{K,\perp}\in\{\Psi^*_K\}^\perp}\frac{(\mathrm{IAA})+(\mathrm{IAB})+(\mathrm{IB})}{\|\mathbb{P}_0^\perp e^{-\mathcal{T}^*(\Pi_K)}\Phi^*_{K,\perp}\|_{\widehat{\mathcal{H}}^1}}\\ &\geqslant \frac{\gamma_K/\omega_K-\varepsilon_{1,K}-\varepsilon_{2,K}}{\big\|\mathbb{P}_0^\perp e^{-\mathcal{T}^*(\Pi_K)}\big\|_{\widehat{\mathcal{H}}^1}}\|e^{-(\mathcal{T}^*)^\dagger}\Phi_{\mathcal{W}}\big\|_{\widehat{\mathcal{H}}^1}\\ &\geqslant \frac{\gamma_K/\omega_K-\varepsilon_{1,K}-\varepsilon_{2,K}}{\|e^{(\mathcal{T}^*)^\dagger}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}\|\mathbb{P}_0^\perp e^{-\mathcal{T}^*(\Pi_K)}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}}\big\|\Phi_{\mathcal{W}}\big\|_{\widehat{\mathcal{H}}^1}.\end{aligned}$$

Defining the constants $\Theta_K := \|e^{(\mathcal{T}^*)^\dagger}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}\|\mathbb{P}_0^\perp e^{-\mathcal{T}^*(\Pi_K)}\|_{\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^1}$, and $\varepsilon_K := \varepsilon_{1,K}+\varepsilon_{2,K}$ and taking the infimum over all coefficient vectors $\boldsymbol{w}_K\in\mathbb{V}_K$ now yields the required estimate. The fact that the constant Θ_K is uniformly bounded above in K is a consequence of the continuity properties of exponential cluster operators together with the density of the union of subspaces $\bigcup_{K\geqslant Q}\mathbb{V}_K$ in $\mathbb{V}$. The fact that the inf-sup constant γ_K is uniformly bounded below in K is a consequence of the eigenvalue convergence $\mathcal{E}^*_K\to\mathcal{E}^*$ (see also the arguments in Remark 2.4.3). □

Equipped with Lemma 2.5.2, we are now ready to state the final result of this section, which concerns the error between the ground state solution of the Full-CC equations in a finite basis (2.40) and the exact solutions of the continuous CC equations (2.17).

Theorem 2.5.5 (Error Estimates for Full-CC in a Finite Basis)**.**

Let the coupled cluster function $f\colon\mathbb{V}\to\mathbb{V}^*$ be defined through Definition 2.4.1, for any $\boldsymbol{t}\in\mathbb{V}$ let $\mathrm{D}f(\boldsymbol{t})$ denote the Fréchet derivative of the coupled cluster function as defined through Equation (2.20), let $\boldsymbol{t}^*\in\mathbb{V}$ denote a zero of the coupled cluster function corresponding to an intermediately normalised eigenfunction $\Psi^*\in\widehat{\mathcal{H}}^1$ of the electronic Hamiltonian $H\colon\widehat{\mathcal{H}}^1\to\widehat{\mathcal{H}}^{-1}$ with isolated, non-degenerate ground state eigenvalue $\mathcal{E}^*$, let $\mathbb{V}_K\subset\mathbb{V}$ denote the Hilbert subspace of coefficients as defined through Definition 2.5.2 for $K\geqslant Q$, let the Full-CI Hamiltonian $H_K\colon\mathcal{V}_K\to\mathcal{V}_K^*$ be defined according to Definition 2.7, for K sufficiently large, let $(\Psi^*_K,\mathcal{E}^*_K)\in\mathcal{V}_K\times\mathbb{R}$ be a sequence of simple eigenpairs of the Full-CI Hamiltonians $\{H_K\}_{K\geqslant Q}$, i.e.,

$$\forall\Phi_K\in\mathcal{V}_K\colon\qquad \langle\Phi_K, H_K\Psi^*_K\rangle_{\mathcal{V}_K\times\mathcal{V}_K^*}=\mathcal{E}^*_K\langle\Phi_K,\Psi^*_K\rangle_{\widehat{\mathcal{H}}^1\times\widehat{\mathcal{H}}^{-1}}\quad \textit{with}\ \ \mathcal{E}^*_K\ \textit{simple}\quad \textit{and such that}$$

$$\lim_{K\to\infty}\|\Psi^*-\Psi^*_K\|_{\widehat{\mathcal{H}}^1}=0,\qquad \lim_{K\to\infty}|\mathcal{E}^*-\mathcal{E}^*_K|=0,$$

and let the constants $\gamma_K,\Theta_K,\varepsilon_K,\omega_K>0$ be defined as in the proof of Lemma 2.5.2.

Then there exists $K_0\in\mathbb{N}$ and a constant $\delta_0>0$ such that for all $K\geqslant K_0$ there exists a unique solution $\boldsymbol{t}^*_K\in\mathbb{V}_K$ to the Full-CC equations in a finite basis (2.40) in the closed ball $\overline{\mathbb{B}_{\delta_K}(\boldsymbol{t}^*)}$ where $\delta_K=\delta_0\dfrac{\gamma_K/\omega_K-\varepsilon_K}{\Theta_K}$.

Moreover, there exists a constant $\mathrm{C}>0$ such that $\forall K\geqslant K_0$ we have the quasi-optimality result

$$\|\boldsymbol{t}^*_K-\boldsymbol{t}^*\|_{\mathbb{V}}\leqslant\mathrm{C}\frac{\Theta_K}{\gamma_K/\omega_K-\varepsilon_K}\inf_{\boldsymbol{s}_K\in\mathbb{V}_K}\|\boldsymbol{s}_K-\boldsymbol{t}^*\|_{\mathbb{V}},\tag{2.53}$$

and we have the residual-based error estimate

$$\|\boldsymbol{t}_K^* - \boldsymbol{t}^*\|_{\mathbb{V}} \leqslant 2 \left\| \mathrm{D}f\left(\boldsymbol{t}_K^*\right)^{-1} \right\|_{\mathbb{V}^* \to \mathbb{V}} \|f\left(\boldsymbol{t}_K^*\right)\|_{\mathbb{V}^*}. \tag{2.54}$$

Proof. As mentioned at the beginning of this section, the Full-Coupled Cluster equations in a finite basis (2.40) are simply a Galerkin discretisation of the continuous coupled cluster equations (2.17). Galerkin discretisations of non-linear equations have been widely studied in the literature on non-linear numerical analysis. In particular, the proof of Theorem 2.5.5 is a direct application of [12, Theorem 7.1]. We merely have to confirm that the assumptions of [12, Theorem 7.1] hold, and this amounts to

1. Establishing that the coupled cluster Fréchet derivative at $\boldsymbol{t}^* \in \mathbb{V}$, which we denote $\mathrm{D}f(\boldsymbol{t}^*)$, satisfies the discrete inf-sup condition

$$\exists \Upsilon_K > 0 \colon \qquad \inf_{0 \neq \boldsymbol{w}_K \in \mathbb{V}_K} \sup_{0 \neq \boldsymbol{s}_K \in \mathbb{V}_K} \frac{\langle \boldsymbol{w}_K, \mathrm{D}f(\boldsymbol{t}^*)\boldsymbol{s}_K \rangle_{\mathbb{V} \times \mathbb{V}^*}}{\|\boldsymbol{w}_K\|_{\mathbb{V}} \|\boldsymbol{w}_S\|_{\mathbb{V}}} \geqslant \Upsilon_K; \tag{2.55}$$

2. Establishing that the coefficient subspaces $\{\mathbb{V}_K\}_{K \geqslant Q}$ satisfy the following approximability condition:

$$\lim_{K \to \infty} \inf_{0 \neq \boldsymbol{s}_K \in \mathbb{V}_K} \frac{1}{\Upsilon_K^2} \|\boldsymbol{t}^* - \boldsymbol{s}_K\|_{\mathbb{V}} = 0. \tag{2.56}$$

The discrete inf-sup condition (2.55) has been established in Lemma 2.5.2 with constant $\Upsilon_K = \dfrac{\gamma_K / \omega_K - \varepsilon_K}{\Theta_K}$ which will obviously be positive for all K sufficiently large since $\varepsilon_K \to 0$. It therefore remains to establish the approximability result (2.56) but this is a simple consequence of the previously exploited fact that the union of subspaces $\bigcup_{K \geqslant Q} \mathbb{V}_K$ is dense in $\mathbb{V}$ together with the fact that, as shown in the proof of Lemma 2.5.2, the constant γ_K is uniformly bounded below in K and the constant Θ_K is uniformly bounded above in K. □

We conclude this section with several remarks.

Remark 2.5.2 (Necessity of Assumptions of Lemma 2.5.2 in Theorem 2.5.5)**.**

Consider the setting of Lemma 2.5.2 and Theorem 2.5.5 and recall in particular the assumption that any isolated, simple eigenpair of the electronic Hamiltonian can be approximated by a sequence of simple eigenpairs of the Full-CI Hamiltonian as expressed through Equation (2.42)*. It is readily seen from the proof of Lemma 2.5.2 that this assumption is not required if one considers invertibility of the CC Fréchet derivative* $\mathrm{D}f(\boldsymbol{t}^*)$ *on* $\mathbb{V}_K$ *at* $\boldsymbol{t}^* = \boldsymbol{t}^*_{\mathrm{GS}}$*. Indeed, in this special case, the discrete inf-sup condition for the shifted Full-CI Hamiltonian* $H_K - \mathcal{E}^*$ *on* $\{\Psi_K^*\}^\perp \subset \mathcal{V}_K$ *can be replaced with the coercivity of the shifted electronic Hamiltonian* $H - \mathcal{E}^*_{\mathrm{GS}}$ *on* $\{\Psi^*_{\mathrm{GS}}\}^\perp \subset \widehat{\mathcal{H}}^1$*. In this special case therefore, the proof of Lemma 2.5.2 holds without any assumption beyond the simplicity of the ground state energy and the intermediate normalisability of the ground state wave-function. Thus we can deduce the asymptotic local well-posedness of the Full-CC equations in a finite-basis* (2.17) *in a neighbourhood of* $\boldsymbol{t}^*_{\mathrm{GS}} \in \mathbb{V}$ *according to Theorem 2.5.5 without any additional assumptions.*

Remark 2.5.3 (Comparing the Conclusions of Theorems 2.5.3 and 2.5.5)**.**

Consider the settings of Theorems 2.5.3 and 2.5.5. Let us emphasise here that, in contrast to Theorem 2.5.3, Theorem 2.5.5 does not explicitly require that the ground state wave-function

in $\mathcal{V}_K$ of the Full-CI Hamiltonian be intermediately normalisable or that the associated ground state eigenvalue be simple. Instead these properties are inherited (for K large enough) from the properties of the exact electronic Hamiltonian $H\colon \widehat{\mathcal{H}}^1 \to \widehat{\mathcal{H}}^{-1}$ thanks to the density of the Q-particle approximation spaces $\{\mathcal{V}_K\}_{K\geqslant Q}$ in $\widehat{\mathcal{H}}^1$. More significantly, Theorem 2.5.5 provides error estimates for the Full-CC equations in a finite basis with respect to the zeros of the exact (infinite-dimensional) coupled cluster function.

Remark 2.5.4 (Error Estimates for the Discrete Coupled Cluster Energies)**.**

It is natural, at this point, to ask whether a priori *and residual-based error estimates of the form* (2.53) *and* (2.54) *can be obtained for the Full-CC discrete energies. Quasi-optimal* a priori *error estimates for the discrete CC energies have been obtained by Schneider and Rohwedder [83, Theorem 4.5] using the dual weighted residual-based approach developed by Rannacher and coworkers [92]. The arguments of Schneider and Rohwedder can readily be seen to apply in our framework, and the proof and statement of [83, Theorem 4.5] can be thus be copied nearly word-for-word, the only difference being that the local monotonicity constant that appears in the* a priori *error estimate in [83, Theorem 4.5] is replaced with the discrete inf-sup constant that we have derived in Lemma 2.5.2. The establishment of residual-based error estimates for the discrete CC energies, which requires considerably more work but can be achieved using the tools developed in this thesis, will be addressed in a forthcoming contribution.*

Part II

Application of a posteriori error estimates for least-cost strategies to achieve target accuracy

Chapter 3

Introduction

The work of Part II was carried out in collaboration with Yvon Maday and Muhammad Hassan.

Numerical methods for solving partial differential equations (PDEs) aim at producing an approximate solution of a given problem. In many cases, such an approximation is obtained without guaranteeing the quality of this solution. The error estimation of such approximate solutions is classified into two types: the *a priori* and *a posteriori* error estimation. The *a priori* error estimation ensures the convergence of the discrete solutions towards the exact one at a certain rate that depends on the degrees of freedom. The *a posteriori* error estimation provides computable error bounds controlling the difference between the approximate result and the true solution. For mesh-based methods, *a posteriori* error estimation also yields the error distribution in different elements and suggest how to refine the mesh in areas with large error. This is called an adaptive mesh refinement (AMR) strategy and it is widely used in finite element method (FEM) [109, 6, 34, 108, 80].

In the past several decades, a large number of articles about the *a posteriori* error estimation have been published, owing to its crucial role in the self-adaptive methods, see e.g., [98, 97, 71, 3, 88] for a more detailed presentation. Our work explores the application of *a posteriori* error estimation for quantum chemistry calculations, which describe the state of basic particles using wave functions and yield the energy of a given through the solution of an eigenvalue problem involving a Hamiltonian. Due to the extremely high-dimensionality of the solution space and the very limited computational resources, approximations are widely used in these numerical calculations. Therefore, *a posteriori* error estimation can serve as a useful and powerful mathematical tool yielding the accuracy of approximate solutions.

Many numerical methods for approximating the state and energy of a chemical system consist of resolving a nonlinear eigenvalue problem. This includes the Gross-Pitaevskii equation which is studied in our work, the Hartree-Fock equations and the Kohn-Sham equations, etc. For the *a priori* error analysis of these equations, we refer to papers [15, 107, 13, 23, 14] and the references therein. The *a posteriori* error analysis for nonlinear eigenvalue problems began with [67] and is followed by [23, 31, 32]. Adaptive refinement methods for nonlinear eigenvalue problem are studied in [32, 31, 21, 22, 54, 101]. Note that these adaptive methods are based on finite element discretisations. For planewave discretisations, we have the *a posteriori* studies in [16, 39] and a first error balance strategy proposed in [38]. The idea of this error balance strategy is to divide the numerical approximation error into two sources which originate respectively from the limited degrees of freedom used in the numerical approximation and the limited number of iterations in

the numerical solution process. The accuracy is improved by decreasing one of those two error source according to their contributions in the total error , which is quantified with a residual-based *a posteriori* error estimator.

The basic strategy in [38] is formulated as follows:

1. The numerical calculation begins by fixing the degree of freedom and increasing the number of iterations until the iteration residual is below the discretisation residual.

2. The degrees of freedom is increased by two.

3. Repeat the above process until the target accuracy is achieved.

Of course, the above stopping criteria and the manner of increasing the degrees of freedom might not be the optimal way of using computational resources. This motivates us to study the optimal strategy of performing numerical calculations in our work.

Under the same context as in [38], the aim of this work is to study how to design numerical calculation strategy such that the simulation is performed efficiently, in the sense that the calculation accuracy and the computational cost are well balanced. Obviously, the accuracy of a numerical approximation can be improved by simply resolving the problem in a large approximation space and adding more iterations, but this is hardly an optimal use of computational resources. In our work, we explore the optimal way to distribute the computation resource for reducing one of the two error sources at all stages of numerical computation. We first explore the optimal (computationally cheapest) way of achieving a given accuracy of the numerical solution using a probabilistic method and then propose near-optimal strategy to achieve given accuracy with a rather low computational cost.

The present work focuses on the numerical solution of Gross-Pitaevskii type equations, a simple but representative problem in quantum chemistry. With the help of a probabilistic method, we first explore the optimal way of performing the numerical calculations, the so-called optimal path. Based on this optimal path, we will summarize some key features of the optimal error balance process. We then aim at proposing near-optimal strategy such that the finally obtained numerical solution achieves target accuracy while maintaining rather cheap calculation cost. The long term goal is to propose near-optimal strategy for the numerical solution of more complex quantum chemistry methods such as the Hartree-Fock equations, Kohn-Sham eqautions, and eventually other ab initio methods.

The remainder of these two chapters is organized as follows. In Chapter 4.1, we introduce a linear elliptic source problem, including the problem description, error analysis and general numerical solution process. In Chapter 4.2, based on these insights, we transfer this linear elliptic source problem to the Gross-Pitaevskii source problem by adding a suitable non-linearity, and we study the solution to this problem. In Chapter 4.3, we introduce the optimal path strategy, which explores the optimal way of numerically solving these above linear and nonlinear problems. Chapter 4.4 offers more complementary tests and analyzes the mechanisms that generates these optimal paths for different problems. In Chapter 4.5, we propose two near-optimal strategies for resolving both the linear and nonlinear problems. In Chapter 5, we apply directly our nearly optimal strategies for the numerical solution of the Gross-Pitaevskii equation to verify the usability of our strategies in solving eigenvalue problem.

Chapter 4

The optimal path problem and near-optimal strategies

4.1 A linear elliptic source problem

4.1.1 Problem description

In this section, we consider the following energy minimization problem

$$E^* = \min \left\{ E(v) := \frac{1}{2}\int_0^{2\pi} |\nabla v|^2 + \frac{1}{2}\int_0^{2\pi} Vv^2 - \langle f, v\rangle_{X',X}\,,\ v \in X \right\}. \tag{4.1}$$

This problem is set on the torus defined as $\mathbb{T} = \mathbb{R}/2\pi\mathbb{Z}$ and $X = H^1(\mathbb{T})$ is the Sobolev space consisting of functions with square-integrable first-order weak derivatives.[1] In addition, we denote by $X' = H^{-1}(\mathbb{T})$ the dual space of X. Besides, we assume that $V \in H^1(\mathbb{T})$, $f \in H^{-1}(\mathbb{T})$ and that V is bounded from below by $V_{\min} > 0$. Any solution $u \in X$ of the above minimization problem (4.1) satisfies the following Euler-Lagrange equation

$$\forall v \in X, \quad \int_0^{2\pi} \nabla u \cdot \nabla v + \int_0^{2\pi} Vuv = \langle f, v\rangle_{X',X}\,. \tag{4.2}$$

The above variational formulation is equivalent to the following representation using duality pairing notation:

$$\forall v \in X, \quad \langle -\Delta u + Vu - f, v\rangle_{X',X} = 0, \tag{4.3}$$

where Δu is defined in the distribution sens and we thus define the linear operator $A := -\Delta + V$ with domain $H^2(\mathbb{T}) \subset X$ and form domain X.

After giving the weak form of this problem, now we show the existence and uniqueness of the solution ,i.e., there exists a unique solution $u \in X$ of the weak problem (4.2), which is, as consequence, also the unique minimizer of the energy functional E defined in (4.1). The above result is a direct consequence by applying the Lax-Milgram Lemma [11] in space X with the bilinear form $a\colon X \times X \to \mathbb{R}$

$$\forall u, v \in X, \quad a(u, v) = \int_0^{2\pi} \nabla u \cdot \nabla v + \int_0^{2\pi} Vuv, \tag{4.4}$$

[1] This space coincides with the space of all functions u in $H^1(0, 2\pi)$ such that $u(0) = u(2\pi)$.

and with the linear functional $b\colon X \to \mathbb{R}$

$$\forall v \in X, \quad b(v) = \langle f, v \rangle_{X',X}\,. \tag{4.5}$$

Indeed, it is easily proven that the bilinear form $a\colon X \times X \to \mathbb{R}$ possesses the continuity and ellipticity properties on X. More precisely,

- The bilinear form $a\colon X \times X \to \mathbb{R}$ defined through Equation (4.4) is continuous on X:

$$\forall v, w \in X, \quad |a(v,w)| \leqslant \beta_a \|v\|_{H^1}\, \|w\|_{H^1}, \tag{4.6}$$

  where $\beta_a = \max\{1, \|V\|_{L^\infty}\}$.

- The bilinear form $a\colon X \times X \to \mathbb{R}$ defined through Equation (4.4) is coercive on X:

$$\forall v \in X, \quad |a(v,v)| \geqslant \gamma_a \|v\|^2_{H^1}, \tag{4.7}$$

  where $\gamma_a = \min\{1, V_{\min}\}$.

Remark 4.1.1.
In the case where f belongs to $L^2(0,2\pi)$, we can also interpret the solution of the minimization problem as follows, it is the solution to the strong problem: find $u \in H^2(0,2\pi)$

$$\begin{cases} -\Delta u + Vu = f, & \text{on } [0,2\pi], \\ u(0) = u(2\pi), u'(0) = u'(2\pi). \end{cases} \tag{4.8}$$

Indeed, we deduce from the regularity of f that $u \in H^2(0,2\pi)$, which guarantees the well-posedness of derivative u'. Next, it is straightforward to deduce the boundary condition $u'(0) = u'(2\pi)$.

After showing the existence and uniqueness of the solution $u \in X$ to the Equation (4.2), now we begin the numerical solution of this problem, which aims at finding a numerical approximate solution in a finite-dimensional subspace $X_N \subset X$ defined below. Applying the Ritz-Galerkin discretisation method over $X_N \subseteq X$ yields a linear system of equations which is resolved using a direct or an iterative method.

In what follows, for the sake of simplicity, we shall restrict our consideration to a specific form of functions f and V, assume them to be even. Consequently, in our case, both v and u are even functions, allowing us to naturally reduce X to its even part:

$$X = H^1_{\text{even}}(\mathbb{T}) := \{v \in H^1(\mathbb{T}) | v \text{ is even function}\}.$$

The set of functions $\{e_i\colon x \mapsto \cos(ix),\ i \in \mathbb{N}^*\} \cup \{e_0\colon x \mapsto \frac{1}{\sqrt{2}}\}$ constitute an orthogonal basis of X. The Fourier spectral approximation of this weak problem consists of defining the family of finite-dimensional subspaces $(X_N)_{N\in\mathbb{N}}$ of X as

$$\forall N \in \mathbb{N}, \quad X_N \coloneqq \text{Span}\{e_i,\ 0 \leqslant i \leqslant N,\ i \in \mathbb{N}\}, \tag{4.9}$$

and looking for discrete solution u_N in the space X_N. Let us recall that for real-valued 2π-periodic even function $v \in X$, we can write

$$\forall x \in [0,2\pi], \quad v(x) = \sum_{i\in\mathbb{N}} \hat{v}_i e_i(x), \tag{4.10}$$

where $\hat{v}_i$ is the i-th Fourier cosine coefficient of v and is defined as

$$\forall i \in \mathbb{N}, \quad \hat{v}_i \coloneqq \frac{1}{\pi}\int_0^{2\pi} v(x)e_i(x)dx.$$

With this choice of basis at hand, we have a simple expression for the norms (L^2 or H^1 or H^{-1}) of v, i.e.,

$$\|v\|_{L^2} = \left(\int_0^{2\pi} v^2\right)^{\frac{1}{2}} = \sqrt{\pi}\left(\sum_{i\in\mathbb{N}} \hat{v}_i^2\right)^{\frac{1}{2}},$$

and for $r \in \{1, -1\}$, we have the function norm defined as follows

$$\|v\|_{H^r} = \sqrt{\pi}\left(\sum_{i\in\mathbb{N}} (1+i^2)^r \hat{v}_i^2\right)^{\frac{1}{2}}. \tag{4.11}$$

More generally, for any $r \in \mathbb{R}$, we define the Sobolev space

$$H^r_{\text{even}}(\mathbb{T}) := \left\{ v\colon \mathbb{T} \to \mathbb{R} \middle| v = \sum_{i\in\mathbb{N}} \hat{v}_i e_i \text{ and } \sum_{i\in\mathbb{N}} (1+i^2)^r \hat{v}_i^2 < +\infty \right\}, \tag{4.12}$$

with corresponding H^r-norm defined as

$$\|v\|_{H^r} = \sqrt{\pi}\left(\sum_{i\in\mathbb{N}} (1+i^2)^r \hat{v}_i^2\right)^{\frac{1}{2}}. \tag{4.13}$$

Remark 4.1.2.

In what follows, we choose

$$\forall x \in [0, 2\pi], \quad V(x) = 1 + \sum_{i\in\mathbb{N}^*} \frac{\cos(ix)}{i^2} \quad \textit{and} \quad f(x) = \sum_{i\in\mathbb{N}^*} \frac{2\cos(ix)}{i^{0.05}}.$$

Note that

$$\forall x \in [0, 2\pi], \quad V(x) = \frac{x^2}{4} - \frac{\pi x}{2} + \frac{\pi^2}{6} + 1,$$

that can be derived, e.g. from [110, P46]. Consequently, it is easy to check that $V \in L^\infty(\mathbb{T}) \cap H^{\frac{3}{2}-\epsilon}$ $(\epsilon > 0)$ *and that the maximum and minimum value of function* V *over* $[0, 2\pi]$ *are respectively* $\|V\|_{L^\infty} = 1 + \frac{\pi^2}{6}$ *and* $V_{\min} = 1 - \frac{\pi^2}{12}$. *Besides,* $f \in H^{-1}(\mathbb{T})$ *as follows from the simple calculation*

$$\|f\|_{H^{-1}} = \left(\sum_{i\in\mathbb{N}^*} \frac{4\pi}{1+i^2}\cdot i^{-0.1}\right)^{\frac{1}{2}} < \left(\sum_{i\in\mathbb{Z}^*} \frac{4\pi}{i^{2.1}}\right)^{\frac{1}{2}} < \infty.$$

After introducing the Fourier spectral approximation, now we state the weak problem in a finite-dimensional discretisation space setting: For $N \in \mathbb{N}^*$, find $u_N \in X_N$ such that

$$\forall v_N \in X_N, \quad a(u_N, v_N) = b(v_N). \tag{4.14}$$

By the Lax-Milgram Lemma, the above discrete problem has exactly one solution. Besides, the

solution u_N also minimizes the energy functional defined through (4.1) over space X_N,

$$E(u_N) = \min_{v_N \in X_N} E(v_N) =: E_N^*. \tag{4.15}$$

4.1.2 *A priori* and *a posteriori* error estimation

The aim of this section is to perform the *a priori* and *a posteriori* analysis. For the *a priori* error analysis, the aim is to guarantee the convergence of the discrete solution $u_N \in X_N$ of problem (4.14) towards the solution $u \in X$ of problem (4.2) when $N \longrightarrow \infty$. The derivation is a direct application of Céa's Lemma [19] and we have the following *a priori* estimation:

$$\|u - u_N\|_{H^1} \leqslant \frac{\beta_a}{\gamma_a} \min_{v_N \in X_N} \|u - v_N\|_{H^1}, \tag{4.16}$$

where β_a is the continuity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.6) and γ_a is the coercivity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.7). Beside, we also have the following energy estimate.

$$\frac{1}{2}\gamma_a \|u - u_N\|_{H^1}^2 \leqslant E(u_N) - E(u) \leqslant \frac{1}{2}\beta_a \|u - u_N\|_{H^1}^2, \tag{4.17}$$

indeed we have: for any $v_N \in X_N$,

$$\begin{aligned} E(v_N) - E(u) &= \frac{1}{2}a(v_N, v_N) - b(v_N) - \frac{1}{2}a(u,u) + b(u) \\ &= \frac{1}{2}a(v_N, v_N) - a(u, v_N) - \frac{1}{2}a(u,u) + a(u,u) \\ &= \frac{1}{2}a(v_N, v_N) - a(u, v_N) + \frac{1}{2}a(u,u) \\ &= \frac{1}{2}a(v_N, v_N) - \frac{1}{2}a(u, v_N) - \frac{1}{2}a(v_N, u) + \frac{1}{2}a(u,u) \\ &= \frac{1}{2}a(v_N - u, v_N - u), \end{aligned} \tag{4.18}$$

where in the above derivation we use the fact that $u \in X$ is the solution of the weak problem (4.2) and that the bilinear form $a\colon X \times X \to \mathbb{R}$ is symmetric. Inserting the discrete solution $u_N \in X_N$ into the above expression and using the continuity and coercivity of $a\colon X \times X \to \mathbb{R}$ in addition, yields the energy estimate given by Equation (4.17).

For any $N \in \mathbb{N}^*$, we define $\Pi_N\colon X \to X_N$ being the L^2-orthogonal projection operator onto X_N and we extend it continuously into X' such that Π_N is defined from X' to X_N. In addition, we define its complementary $\Pi_N^\perp \coloneqq \mathbf{I} - \Pi_N$. Then, for any $v \in X$ with Fourier expansion

$$v = \sum_{i \in \mathbb{N}} \hat{v}_i e_i,$$

we have

$$\Pi_N v = \sum_{0 \leqslant i \leqslant N} \hat{v}_i e_i \text{ and } \Pi_N^\perp v = \sum_{i > N} \hat{v}_i e_i.$$

The convergence rate of the truncated series $\Pi_N v$ towards v depends on the regularity of function v: for any real numbers r and s with $s < r$, we have [18]

$$\forall v \in H^r(\mathbb{T}), \quad \|v - \Pi_N v\|_{H^s} \leqslant \frac{1}{N^{r-s}} \|v - \Pi_N v\|_{H^r} \leqslant \frac{1}{N^{r-s}} \|v\|_{H^r}. \tag{4.19}$$

Moreover, we point out that the L^2-orthogonal projection operator Π_N also satisfies the following H^1-orthogonal relation

$$\forall v \in X,\ \forall w_N \in X_N, \quad \int_0^{2\pi} \nabla(v - \Pi_N v) \cdot \nabla w_N + \int_0^{2\pi} (v - \Pi_N v) w_N = 0. \tag{4.20}$$

In addition to the above classical *a priori* error estimate given by Equation (4.16), we also have the following classical L^2 error estimate based on the Aubin-Nitsche duality argument (see, e.g., [9, Theorem 5.4.8]).

Lemma 4.1.1. *Let $u \in X$ be the weak solution of problem (4.2) and for $N \in \mathbb{N}^*$, let $u_N \in X_N$ be the solution of discrete problem (4.14). Then there exits constant $C_1 > 0$ such that*

$$\|u - u_N\|_{L^2} \leqslant C_1 N^{-1} \|u - u_N\|_{H^1}, \tag{4.21}$$

Remark 4.1.3.

In the following chapters of the thesis, by an abuse of notation, we will use the constant C_1, C_2 or C_3 several times in different lemmas, theorems. It is important to note that in each statement, those constants represent different values and each specific value will be defined within the proof.

Proof. For showing the L^2 convergence, we firstly introduce the following adjoint problem: find $\varphi \in X$ such that

$$\forall v \in X, \quad a(v, \varphi) = \int_0^{2\pi} v(u - u_N), \tag{4.22}$$

where a is the bilinear form defined in the weak problem (4.2).

The existence and uniqueness of the solution φ is a direct consequence by applying the Lax-Milgram Lemma. In addition, we have the following classical elliptic regularity result [40]: there exists constant $c_1 > 0$ such that

$$\|\varphi\|_{H^2} \leqslant \frac{1}{c_1} \|u - u_N\|_{L^2}. \tag{4.23}$$

Combining the above estimate with Estimate (4.19) and recalling the Galerkin orthogonality that we have $a(u - u_N, v_N) = 0$ for any $v_N \in X_N$, we deduce that

$$\begin{aligned}
\|u - u_N\|_{L^2}^2 &= \int_0^{2\pi} (u - u_N)^2 \\
&= a(\varphi, u - u_N) \\
&= a(\varphi - \Pi_N \varphi, u - u_N) \\
&\leqslant \beta_a \|\varphi - \Pi_N \varphi\|_{H^1}\ \|u - u_N\|_{H^1} \\
&\leqslant \beta_a N^{-1} \|\varphi\|_{H^2}\ \|u - u_N\|_{H^1} \\
&\leqslant \frac{\beta_a}{c_1 N} \|u - u_N\|_{L^2}\ \|u - u_N\|_{H^1},
\end{aligned}$$

where β_a is the continuity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.6). From the above derivation we deduce directly (4.21) and we have $C_1 = \frac{\beta_a}{c_1}$. □

Additionally, we have the following super convergence result.

Theorem 4.1.1. *Let $u \in X$ be the weak solution of problem (4.2), for $N \in \mathbb{N}^*$, let $u_N \in X_N$ be the solution of discrete problem (4.14), let $\Pi_N \colon X' \to X_N$ be the orthogonal projection operator onto space X_N, let $V \in H^s(\mathbb{T})$ for some $s > \frac{1}{2}$ be the potential function in problem (4.2) and let us denote by $r = \min\{2, s\}$. Then for N large enough:*

- *There exits constant $C_2 > 0$ such that*

$$\|\Pi_N u - u_N\|_{H^1} \leqslant C_2 N^{-\frac{r}{2}-1} \|u - u_N\|_{H^1}. \tag{4.24}$$

- *There exits constant $C_3 > 0$ such that*

$$\|\Pi_N u - u_N\|_{L^2} \leqslant C_3 N^{-r-1} \|u - u_N\|_{H^1}. \tag{4.25}$$

Specifically, for our chosen potential function V defined in Remark 4.1.2, we have

$$\|\Pi_N u - u_N\|_{H^1} \leqslant C_2 N^{-\frac{7}{4}+\epsilon} \|u - u_N\|_{H^1} \tag{4.26}$$

and

$$\|\Pi_N u - u_N\|_{L^2} \leqslant C_3 N^{-\frac{5}{2}+\epsilon} \|u - u_N\|_{H^1}, \tag{4.27}$$

where $\epsilon > 0$.

Proof. At the first step, using the Galerkin orthogonality relation, we have the following equality:

$$a(u_N - \Pi_N u, u_N - \Pi_N u) = a(u - \Pi_N u, u_N - \Pi_N u). \tag{4.28}$$

For the left-hand side of Equation (4.28), we have

$$a(u_N - \Pi_N u, u_N - \Pi_N u) \geqslant \gamma_a \|u_N - \Pi_N u\|^2_{H^1}, \tag{4.29}$$

where γ_a is the coercivity constant of $a \colon X \times X \to \mathbb{R}$ defined through (4.7). For the right-hand side of Equation (4.28), using the orthogonality relation (4.20), we deduce that

$$\begin{aligned} a(u - \Pi_N u, u_N - \Pi_N u) &= \int_0^{2\pi} \nabla(u - \Pi_N u) \cdot \nabla(u_N - \Pi_N u) + \int_0^{2\pi} V(u - \Pi_N u)(u_N - \Pi_N u) \\ &= \int_0^{2\pi} V(u - \Pi_N u)(u_N - \Pi_N u) \\ &\leqslant \|V\|_{L^\infty} \ \|u - \Pi_N u\|_{L^2} \ \|u_N - \Pi_N u\|_{L^2}. \end{aligned} \tag{4.30}$$

For the aim of bounding the term $\|u_N - \Pi_N u\|_{L^2}$, we introduce a new adjoint problem: find $\vartheta \in X$ such that

$$\forall v \in X, \quad a(v, \vartheta) = \int_0^{2\pi} v(u_N - \Pi_N u). \tag{4.31}$$

Similarly to the problem (4.22), the existence and uniqueness of the solution ϑ is a direct consequence by applying the Lax-Milgram Lemma and there exists $c_2 > 0$ such that

$$\|\vartheta\|_{H^2} \leqslant c_2 \|u_N - \Pi_N u\|_{L^2}. \tag{4.32}$$

And we therefore have

$$\begin{aligned}\|u_N - \Pi_N u\|_{L^2}^2 &= \int_0^{2\pi} (u_N - \Pi_N u)^2 \\ &= a(\vartheta, u_N - \Pi_N u) \\ &= a(\vartheta - \Pi_N \vartheta, u_N - \Pi_N u) + a(\Pi_N \vartheta, u_N - \Pi_N u) \\ &= a(\vartheta - \Pi_N \vartheta, u_N - \Pi_N u) + a(\Pi_N \vartheta, u - \Pi_N u)\end{aligned}$$

where the last step comes from the Galerkin orthogonality. Recalling Equation (4.20) and Estimate (4.19), we deduce that

$$\begin{aligned}\|u_N - \Pi_N u\|_{L^2}^2 &= a(\vartheta - \Pi_N \vartheta, u_N - \Pi_N u) + \int_0^{2\pi} \nabla(\Pi_N \vartheta) \cdot \nabla(u - \Pi_N u) + \int_0^{2\pi} V \Pi_N \vartheta (u - \Pi_N u) \\ &= a(\vartheta - \Pi_N \vartheta, u_N - \Pi_N u) + \int_0^{2\pi} V \Pi_N \vartheta (u - \Pi_N u) - \int_0^{2\pi} \Pi_N (V \Pi_N \vartheta)(u - \Pi_N u) \\ &= a(\vartheta - \Pi_N \vartheta, u_N - \Pi_N u) + \int_0^{2\pi} [V \Pi_N \vartheta - \Pi_N (V \Pi_N \vartheta)](u - \Pi_N u) \\ &\leqslant \|V\|_{L^\infty} \, \|\vartheta - \Pi_N \vartheta\|_{L^2} \, \|u_N - \Pi_N u\|_{L^2} + \|V \Pi_N \vartheta - \Pi_N (V \Pi_N \vartheta)\|_{L^2} \, \|u - \Pi_N u\|_{L^2} \\ &\leqslant N^{-2} \|V\|_{L^\infty} \, \|\vartheta\|_{H^2} \, \|u_N - \Pi_N u\|_{L^2} + N^{-r} \|V \Pi_N \vartheta\|_{H^r} \, \|u - \Pi_N u\|_{L^2} \\ &\leqslant N^{-2} \|V\|_{L^\infty} \, \|\vartheta\|_{H^2} \, \|u_N - \Pi_N u\|_{L^2} + N^{-r} \|V\|_{H^r} \, \|\Pi_N \vartheta\|_{H^r} \, \|u - \Pi_N u\|_{L^2} \\ &\leqslant c_2 N^{-2} \|V\|_{L^\infty} \, \|u_N - \Pi_N u\|_{L^2}^2 + c_2 N^{-r} \|V\|_{H^r} \, \|u - \Pi_N u\|_{L^2} \, \|u_N - \Pi_N u\|_{L^2},\end{aligned}$$

where the last step comes from Estimate (4.32). For N large enough such that $c_2 N^{-2} \|V\|_{L^\infty} \leqslant \frac{1}{2}$, it therefore follows that

$$\begin{aligned}\frac{1}{2} \|u_N - \Pi_N u\|_{L^2}^2 &\leqslant c_2 N^{-r} \|V\|_{H^r} \, \|u - \Pi_N u\|_{L^2} \, \|u_N - \Pi_N u\|_{L^2} \\ \|u_N - \Pi_N u\|_{L^2} &\leqslant 2 c_2 N^{-r} \|u - \Pi_N u\|_{L^2}.\end{aligned} \tag{4.33}$$

Combining Estimates (4.19), (4.29), (4.30) and (4.33) yields that

$$\begin{aligned}\frac{\gamma_a}{\|V\|_{L^\infty}} \|u_N - \Pi_N u\|_{H^1}^2 &\leqslant 2 c_2 N^{-r} \|u - \Pi_N u\|_{L^2}^2 \\ &\leqslant 2 c_2 N^{-r-2} \|u - \Pi_N u\|_{H^1}^2,\end{aligned}$$

from which we deduce that

$$\|u_N - \Pi_N u\|_{H^1} \leqslant c_3 N^{-\frac{r}{2}-1} \left(\|u - u_N\|_{H^1} + \|u_N - \Pi_N u\|_{H^1} \right), \tag{4.34}$$

where $c_3 = \left(\frac{2 c_2 \|V\|_{L^\infty}}{\gamma_a} \right)^{\frac{1}{2}}$. For N large enough such that $c_3 N^{-\frac{r}{2}-1} \leqslant \frac{1}{2}$, we have

$$\|u_N - \Pi_N u\|_{H^1} \leqslant 2 c_3 N^{-\frac{r}{2}-1} \|u - u_N\|_{H^1}, \tag{4.35}$$

which demonstrates the Estimate (4.26).

In addition, for N large enough such that $2 c_2 N^{-r} \leqslant \frac{1}{2}$, combining Estimates (4.21) and

(4.33) yields that

$$\|u_N - \Pi_N u\|_{L^2} \leqslant 4c_2 N^{-r-1} \|u - u_N\|_{H^1},$$

which demonstrates the convergence result (4.27). □

In the first part of this section, we perform the *a priori* error analysis of this problem. The *a priori* analysis bounds the term $\|u - u_N\|_{H^1}$ by $\min_{v_N \in X_N} \|u - v_N\|_{H^1}$, which guarantees the convergence of the discrete solution u_N towards the exact solution u, assuming that u_N can be computed. In the rest of this section, we will derive the *a posteriori* error analysis. The *a posteriori* error analysis aims at providing a bound of $\|u - \widetilde{u_N}\|_{H^1}$ where $\widetilde{u_N}$ is an available numerical approximation of the discrete solution u_N. In contrast to the *a priori* analysis, the *a posteriori* bound is fully computable and (hopefully) should be cheap to calculate. It is a function of the approximation solution $\widetilde{u_N}$ and other computable parameter. There are various strategies for deriving *a posteriori* error estimation [96]. In this work, we use the residual-based *a posteriori* error estimator and the residual is defined as

$$R(\widetilde{u_N}) \coloneqq -\Delta \widetilde{u_N} + V \widetilde{u_N} - f. \tag{4.36}$$

Here we state the lemma showing that the error is bounded by the residual under a suitable norm, which is based on the classical estimate for elliptic partial differential equation (see, e.g., [99]).

Proposition 4.1.1 (*A posteriori* error estimation)**.** *Let $u \in X$ be the weak solution of problem* (4.2)*, for $N \in \mathbb{N}^*$, let X_N be the discrete space defined through* (4.9)*, let β_a be the continuity constant of $a\colon X \times X \to \mathbb{R}$ defined through* (4.6)*, let γ_a be the coercivity constant of $a\colon X \times X \to \mathbb{R}$ defined through* (4.7) *and let $E\colon X \to \mathbb{R}$ be the energy functional defined though* (4.1)*. Then, for any $\widetilde{u_N} \in X_N$ we have*

$$\|\widetilde{u_N} - u\|_{H^1} \leqslant \frac{1}{\gamma_a} \|R(\widetilde{u_N})\|_{H^{-1}}, \tag{4.37}$$

and

$$0 \leqslant E(\widetilde{u_N}) - E(u) \leqslant \frac{\beta_a}{2\gamma_a^2} \|R(\widetilde{u_N})\|_{H^{-1}}^2. \tag{4.38}$$

Proof. Recalling the bilinear form $a\colon X \times X \to \mathbb{R}$ and the functional $b\colon X \to \mathbb{R}$ defined through (4.4) and (4.5) respectively, we deduce that

$$\begin{aligned}
a(\widetilde{u_N} - u, \widetilde{u_N} - u) &= a(\widetilde{u_N}, \widetilde{u_N} - u) - a(u, \widetilde{u_N} - u) \\
&= a(\widetilde{u_N}, \widetilde{u_N} - u) - b(\widetilde{u_N} - u) - a(u, \widetilde{u_N} - u) + b(\widetilde{u_N} - u) \\
&= a(\widetilde{u_N}, \widetilde{u_N} - u) - b(\widetilde{u_N} - u) \\
&= \langle -\Delta \widetilde{u_N} + V \widetilde{u_N} - f, \widetilde{u_N} - u \rangle_{X',X} \\
&\leqslant \| - \Delta \widetilde{u_N} + V \widetilde{u_N} - f\|_{H^{-1}} \|\widetilde{u_N} - u\|_{H^1},
\end{aligned}$$

where in the above expression we make use of the fact that $u \in X$ is the solution of the weak problem (4.2). Together with the coercivity property of the bilinear form $a\colon X \times X \to \mathbb{R}$, we deduce that

$$\|\widetilde{u_N} - u\|_{H^1} \leqslant \frac{1}{\gamma_a} \| - \Delta \widetilde{u_N} + V \widetilde{u_N} - f\|_{H^{-1}}.$$

For the estimation of the energy error, we make use of the Equation (4.18). Indeed, we have

$$E(\widetilde{u_N}) - E(u) = \frac{1}{2}a(\widetilde{u_N} - u, \widetilde{u_N} - u).$$

Equipped with the continuity constant β_a of $a\colon X \times X \to \mathbb{R}$ defined through (4.6), we deduce form the solution error estimate that

$$E(\widetilde{u_N}) - E(u) \leqslant \frac{\beta_a}{2\gamma_a^2} \| - \Delta\widetilde{u_N} + V\widetilde{u_N} - f\|^2_{H^{-1}}.$$

□

4.1.3 Iteration scheme and analysis

In this section, we give the details of numerical solution of this problem. More importantly, in this section, we introduce the concepts of different error source and following this error balance idea, we will present the optimal path problem in Chapter 4.3.

Based on the discretized weak problem (4.14), by inserting different test functions $\{e_j\}_{0\leqslant j\leqslant N}$ (vectors of the basis chosen in (4.9)), we obtain system of linear equations written in the following matrix form

$$\mathbf{A}_N\mathbf{u}_N = \mathbf{b}_N, \tag{4.39}$$

where $\mathbf{A}_N = [a(e_i, e_j)]_{0\leqslant i,j\leqslant N}$, $\mathbf{b}_N = [b(e_j)]_{0\leqslant j\leqslant N}$ and $\mathbf{u}_N$ is the unknown vector of coefficients $\mathbf{u}_N = [\widetilde{u_{N0}}, \widetilde{u_{N1}}, \ldots, \widetilde{u_{NN}}]^T$. Most of the times, the above matrix equation is resolved using iterative method. In this work, we use the Gauss-Seidel-Relaxation(GSR) method. Starting from an initial guess $\mathbf{u}_N^0$, in each iteration step the new vector $\mathbf{u}_N^{k+1}$ is calculated from $\mathbf{u}_N^k$ via the following expression:

$$\mathbf{u}_N^{k+1} = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N\right)\mathbf{u}_N^k + (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N, \tag{4.40}$$

where ω is the relaxation factor, $\mathbf{U}_N$, $\mathbf{D}_N$ and $\mathbf{L}_N$ are respectively the strict upper triangular component, the diagonal component and the strict lower triangular component of matrix $\mathbf{A}_N$. Here, we remark that $\mathbf{U}_N = \mathbf{L}_N^T$ which will be used in the iterative scheme convergence analysis. By defining $\mathbf{P}_N := -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N + \mathbf{U}_N)$ and $\mathbf{q}_N := (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N$, the above expression can be simplified as the following more general form

$$\mathbf{u}_N^{k+1} = \mathbf{P}_N\mathbf{u}_N^k + \mathbf{q}_N. \tag{4.41}$$

From the iterative scheme given by Equation (4.41), we know that the converged solution $\mathbf{u}_N^\infty$ also satisfies this equality

$$\mathbf{u}_N^\infty = \mathbf{P}_N\mathbf{u}_N^\infty + \mathbf{q}_N. \tag{4.42}$$

Subtracting Equation (4.41) by (4.42) yields that

$$\mathbf{u}_N^{k+1} - \mathbf{u}_N^\infty = \mathbf{P}_N(\mathbf{u}_N^k - \mathbf{u}_N^\infty). \tag{4.43}$$

Denoting the iteration error by $\mathbf{e}_N^k := \mathbf{u}_N^k - \mathbf{u}_N^\infty$ yields the following expression which is similar to the power iteration scheme:

$$\mathbf{e}_N^{k+1} = \mathbf{P}_N\mathbf{e}_N^k. \tag{4.44}$$

According to the analysis with the power iteration scheme, it follows that the iteration error $\mathbf{e}_N^k$ goes to $\mathbf{0}$ as long as the spectral radius of $\mathbf{P}_N$ is strictly smaller than 1: $\rho(\mathbf{P}_N) < 1$, where the

spectral radius is defined as maximal of absolute value of all eigenvalues of the square matrix $\mathbf{P}_N$. We denote by $\mathbf{v}_{\max}^{\mathbf{P}}$ the eigenvector of $\mathbf{P}_N$ corresponding to the biggest eigenvalue (in absolute value) which we denote by $\lambda_{\max}^{\mathbf{P}}$. Supposing that the initial error $\mathbf{e}_N^0$ is decomposed as linear combination of eigenvectors of the matrix $\mathbf{P}_N$ and that the coefficient of $\mathbf{v}_{\max}^{\mathbf{P}}$ is non-zero. After a large number of iterations, the iteration error $\mathbf{e}_N^k$ will behave to be proportional to $\mathbf{v}_{\max}^{\mathbf{P}}$ and the ratio of iteration error between two successive iterations $\frac{\mathbf{e}_N^{k+1}}{\mathbf{e}_N^k}$ will be smaller than $\lambda_{\max}^{\mathbf{P}}$ and very close to $\lambda_{\max}^{\mathbf{P}}$. Similarly, the above statement holds for iteration increment denoted by $\widetilde{\mathbf{e}_N^k} := \mathbf{u}_N^{k+1} - \mathbf{u}_N^k$ and this will be checked numerically in next section. Without the restriction of large number of iterations, we have the following more general expression.

$$\|\mathbf{u}_N^{k+1} - \mathbf{u}_N^{\infty}\|_{\ell^2} \leqslant \|\mathbf{P}_N\|_{\ell^2 \to \ell^2} \, \|\mathbf{u}_N^k - \mathbf{u}_N^{\infty}\|_{\ell^2}, \tag{4.45}$$

where $\|\cdot\|_{\ell^2}$ is the Euclidean vector norm. Note that for any function $v \in X$ with corresponding Fourier coefficients vector $\mathbf{v} = [\widehat{v}_j]_{0 \leqslant j \leqslant N}$, we have the following relation:

$$\|\mathbf{v}\|_{L^2}^2 = \mathbf{v}^T \mathbf{v} = \sum_{j=0}^{N} \hat{v}_j^2 = \frac{1}{\pi} \|v\|_{L^2}^2.$$

Thus, for the sake of simplicity, we also refer $\|\cdot\|_{L^2}$ to the ℓ^2 Euclidean vector norm. In analogue to the above relation, we define the H^1 vector norm in a similar way such that for any function $v \in X$ with corresponding Fourier coefficients vector $\mathbf{v} = [\widehat{v}_j]_{0 \leqslant j \leqslant N}$, we have

$$\|\mathbf{v}\|_{H^1}^2 \coloneqq (\mathbf{v}_{H^1})^T \mathbf{v}_{H^1} = \sum_{j=0}^{N} (1 + j^2) \hat{v}_j^2 = \frac{1}{\pi} \|v\|_{H^1}^2.$$

Therefore, we define the H^1 (by an abuse of notation) vector norm as

$$\|\mathbf{v}\|_{H^1} := \big((\mathbf{T}_N \mathbf{v})^T (\mathbf{T}_N \mathbf{v})\big)^{\frac{1}{2}} = \|\mathbf{T}_N \mathbf{v}\|_{L^2}, \tag{4.46}$$

where the matrix $\mathbf{T}_N$ is a diagonal matrix with the diagonal part $[(1 + j^2)^{\frac{1}{2}}]_{0 \leqslant j \leqslant N}$. In addition, the matrix $\mathbf{T}_N$ is invertible and its inverse $\mathbf{T}_N^{-1}$ is also a diagonal matrix with the diagonal part $[(1 + j^2)^{-\frac{1}{2}}]_{0 \leqslant j \leqslant N}$.

In analogue to Estimate (4.45), we have the following estimate

$$\begin{aligned} \|\mathbf{u}_N^{k+1} - \mathbf{u}_N^{\infty}\|_{H^1} &= \|\mathbf{T}_N \mathbf{e}_N^{k+1}\|_{L^2} \\ &= \|\mathbf{T}_N \mathbf{P}_N \mathbf{e}_N^k\|_{L^2} \\ &= \|\mathbf{T}_N \mathbf{P}_N \mathbf{T}_N^{-1} \mathbf{T}_N \mathbf{e}_N^k\|_{L^2} \\ &\leqslant \|\mathbf{T}_N \mathbf{P}_N \mathbf{T}_N^{-1}\|_{L^2 \to L^2} \, \|\mathbf{T}_N \mathbf{e}_N^k\|_{L^2} \\ &= \|\mathbf{T}_N \mathbf{P}_N \mathbf{T}_N^{-1}\|_{L^2 \to L^2} \, \|\mathbf{u}_N^k - \mathbf{u}_N^{\infty}\|_{H^1}, \end{aligned} \tag{4.47}$$

from which we define the matrix norm $\|\mathbf{P}_N\|_{H^1 \to H^1}$ as

$$\|\mathbf{P}_N\|_{H^1 \to H^1} = \sqrt{\lambda_{\max}\left((\mathbf{T}_N \mathbf{P}_N \mathbf{T}_N^{-1})^T (\mathbf{T}_N \mathbf{P}_N \mathbf{T}_N^{-1})\right)} = \sigma_{\max}(\mathbf{T}_N \mathbf{P}_N \mathbf{T}_N^{-1}).$$

Here we lastly remark that the matrix norm $\|\mathbf{P}_N\|_{H^1 \to H^1}$ is the sharper upper bound of the

error decrease ratio $\frac{\|\mathbf{u}_N^{k+1}-\mathbf{u}_N^{\infty}\|_{H^1}}{\|\mathbf{u}_N^{k}-\mathbf{u}_N^{\infty}\|_{H^1}}$ and it is possible that the true ratio is smaller, which means that the error decreases faster.

Recall that the iterative scheme given by Equation (4.40) for resolving this matrix equation can be rewritten as follows

$$(\mathbf{D}_N + \omega\mathbf{L}_N)\mathbf{u}_N^k + ((1-\omega)\mathbf{L}_N + \mathbf{U}_N)\,\mathbf{u}_N^{k-1} = \mathbf{b}_N. \tag{4.48}$$

And the above expression can be simplified as

$$\mathbf{A}_{1,N}\mathbf{u}_N^k + \mathbf{A}_{2,N}\mathbf{u}_N^{k-1} = \mathbf{b}_N. \tag{4.49}$$

where $\mathbf{A}_{1,N} \coloneqq \mathbf{D}_N + \omega\mathbf{L}_N$ and $\mathbf{A}_{2,N} \coloneqq (1-\omega)\mathbf{L}_N + \mathbf{U}_N$. As the above matrix decomposition of $\mathbf{A}_N$ in the GSR scheme, we introduce the following operator decomposition defined in the continuous space X, corresponding to the discrete matrix decomposition of $\mathbf{A}_N$.

Proposition 4.1.2. *Let a be the bilinear form defined in (4.4) and let $(e_i)_{i\in\mathbb{N}}$ be the L^2-orthogonal basis of X chosen in (4.9). Then for any $v \in X$ written in the form of cosine series $v = \sum_{j\in\mathbb{N}} \hat{v}_j e_j$,*

- *we define the operator $A\colon X \to X'$ as the mapping with the property that*

$$Av = -\Delta v + Vv. \tag{4.50}$$

- *we define the operator $D\colon X \to X'$ as the mapping with the property that*

$$Dv = \sum_{j\in\mathbb{N}} a(e_j, e_j)\hat{v}_j e_j, \tag{4.51}$$

- *we define the operator $L\colon X \to X'$ as the mapping with the property that*

$$Lv = \sum_{i\in\mathbb{N}}\sum_{j<i} a(e_i, e_j)\hat{v}_j e_i, \tag{4.52}$$

- *we define the operator $U\colon X \to X'$ as the mapping with the property that*

$$Uv = \sum_{i\in\mathbb{N}}\sum_{j>i} a(e_i, e_j)\hat{v}_j e_i. \tag{4.53}$$

Proof. The proof aims at showing the well-posedness of the above four operators, i.e., showing that they are mappings defined from X to X'.

Let us begin by showing the well-posedness of operator A, which is a direct consequence of the continuity of bilinear form a. In fact, for any $v, w \in X$, we have

$$\langle Av, w\rangle_{X',X} = a(v,w) \leqslant \beta_a \|v\|_{H^1}\ \|w\|_{H^1},$$

where β_a is the continuity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.6). And from the above derivation, we deduce immediately that

$$\|Av\|_{H^{-1}} = \sup_{w\in X, w\neq 0} \frac{\langle Av, w\rangle_{X',X}}{\|w\|_{H^1}} \leqslant \beta_a \|v\|_{H^1} \leqslant \infty, \tag{4.54}$$

which proofs that $A\colon X \to X'$ is well defined.

Now we focus on the diagonal operator. Recall the expression of H^{-1} norm defined in (4.11), for any $v = \sum_{j\in\mathbb{N}} \hat{v}_j e_j \in X$, we have

$$\begin{aligned}
\|Dv\|_{H^{-1}} &= \sqrt{\pi}\left(\sum_{j\in\mathbb{N}} \frac{a(e_j,e_j)^2\hat{v}_j^2}{1+j^2}\right)^{\frac{1}{2}} \\
&\leqslant \sqrt{\pi}\left(\sum_{j\in\mathbb{N}} \frac{\beta_a^2\|e_j\|_{H^1}^4\hat{v}_j^2}{1+j^2}\right)^{\frac{1}{2}} \\
&\leqslant \pi\beta_a\left(\sum_{j\in\mathbb{N}} \frac{(1+j^2)^2\hat{v}_j^2}{1+j^2}\right)^{\frac{1}{2}} \\
&\leqslant \sqrt{\pi}\beta_a\|v\|_{H^1},
\end{aligned}$$

where β_a is the continuity constant of $a\colon X\times X\to\mathbb{R}$ defined through (4.6).

$$\|Dv\|_{H^{-1}} \leqslant \sqrt{\pi}\beta_a\|v\|_{H^1} < \infty, \tag{4.55}$$

which shows the well-posedness of operator $D\colon X\to X'$. Next, we show the well-posedness of operator $L\colon X\to X'$. For any $v=\sum_{j\in\mathbb{N}}\hat{v}_j e_j\in X$, we have

$$\|Lv\|_{H^{-1}} = \sqrt{\pi}\left(\sum_{i\in\mathbb{N}} \frac{\left(\sum_{j<i} a(e_i,e_j)\hat{v}_j\right)^2}{1+i^2}\right)^{\frac{1}{2}}. \tag{4.56}$$

For any $i>j\geqslant 0$, we have

$$a(e_i,e_j) = \int_0^{2\pi} Ve_ie_j \leqslant \|V\|_{L^\infty}\ \|e_i\|_{L^2}\ \|e_j\|_{L^2} \leqslant \pi\|V\|_{L^\infty}. \tag{4.57}$$

Inserting Estimate (4.57) into (4.56) yields

$$\begin{aligned}
\|Lv\|_{H^{-1}} &\leqslant \pi^{\frac{3}{2}}\|V\|_{L^\infty}\left(\sum_{i\in\mathbb{N}}\frac{\left(\sum_{j<i}|\hat{v}_j|\right)^2}{1+i^2}\right)^{\frac{1}{2}} \\
&\leqslant \pi^{\frac{3}{2}}\|V\|_{L^\infty}\left(\sum_{i\in\mathbb{N}}\frac{\left(\sum_{j<i}\frac{1}{1+j^2}\right)\left(\sum_{j<i}(1+j^2)\hat{v}_j^2\right)}{1+i^2}\right)^{\frac{1}{2}} \\
&\leqslant \pi^{\frac{3}{2}}\|V\|_{L^\infty}\left(\sum_{i\in\mathbb{N}}\frac{\left(\sum_{j\in\mathbb{N}}\frac{1}{1+j^2}\right)\left(\sum_{j\in\mathbb{N}}(1+j^2)\hat{v}_j^2\right)}{1+i^2}\right)^{\frac{1}{2}} \\
&\leqslant \pi^2\|V\|_{L^\infty}\ \|v\|_{H^1}\left(\sum_{i\in\mathbb{N}}\frac{1}{1+i^2}\right)^{\frac{1}{2}} \\
&\leqslant \infty,
\end{aligned} \tag{4.58}$$

where we have made use of the Cauchy-Schwarz Inequality in the derivation.

The above inequality shows that L defines a mapping from X to its dual space X'. In the end, we show the well-posedness of operator U, whose proof is quite similar as the one used for operator L. For $i \in \mathbb{N}$, in both cases where $j < i$ or $j > i$, the partial sum of $\hat{v}_j^2$ is always bounded by $\|v\|_{H^1}^2 \sum_{j\in\mathbb{N}} \frac{1}{1+j^2}$. Therefore, we deduce that the mapping $U\colon X \to X'$ is well defined, which completes the proof. □

Corollary 4.1.1.1. *Let operator D, L and U be defined through Proposition 4.1.2 and let $\omega \in \mathbb{R}$ be the relaxation parameter of the GSR iteration scheme. Then we define operator $A_1 \coloneqq D + \omega L$ and operator $A_2 \coloneqq (1-\omega)L + U$ satisfying $A = A_1 + A_2$, which corresponds to the matrix decomposition $\mathbf{A}_N = \mathbf{A}_{1,N} + \mathbf{A}_{2,N}$ in GSR scheme.*

Recall that for any $N \in \mathbb{N}^*$, $\Pi_N\colon X' \to X_N$ is the extended L^2-orthogonal projection operator onto space X_N. Here, based on the relation between function $u_N \in X_N$ and its Fourier coefficient vector $\mathbf{u}_N$, by an abuse of notation, we also refer Π_N to the vector projection operator picking the first $N+1$ elements in the coefficient vector $\mathbf{u}_M$ corresponding to $u_M \in X_M (M > N)$. Combining with Corollary 4.1.1.1, we deduce immediately that for any $N \in \mathbb{N}^*$, we have

$$\Pi_N A_1 u_N^k + \Pi_N A_2 u_N^{k-1} = \Pi_N f. \tag{4.59}$$

In addition, we have the following discretized version of equivalence:

$$u_N \in X_N \text{ is solution of the weak problem (4.14)} \quad \Longleftrightarrow \quad u_N \in X_N,\ \Pi_N A(u_N) = \Pi_N f. \tag{4.60}$$

Moreover, for any $N \in \mathbb{N}$, by denoting $A_{1,N} \coloneqq \Pi_N A_1 \Pi_N$, we have $A_{1,N} = \Pi_N A_1 \Pi_N = \Pi_N(\Pi_M A_1 \Pi_M)\Pi_N = \Pi_N A_{1,M} \Pi_N$ and this relation also holds if we replace A_1 by A_2 or A.

The error between the exact solution u and our approximate solution u_N^k obtained after k iterations stems mainly from two sources: the finite number of iterations and the finite-dimensional discretisation space. According to this idea, we split the total error into iteration error $\mathrm{Er}_{\mathrm{iter}}$ and

discretisation error $\mathrm{Er}_{\mathrm{disc}}$ in two different manners:

$$\|u_N^k - u\|_{H^1} \leqslant \underbrace{\|u_N^k - u_N\|_{H^1}}_{:=\mathrm{Er}_{\mathrm{iter}}(u_N^k)} + \underbrace{\|u_N - u\|_{H^1}}_{:=\mathrm{Er}_{\mathrm{disc}}^N}, \tag{4.61}$$

where u_N is the solution of discrete problem (4.14). In this definition, the term $u_N - u$ represents purely the discretisation error without influence of finite number of iterations. From another point of view, we define these two source errors as:

$$\|u_N^k - u\|_{H^1} \leqslant \underbrace{\|u_N^k - u^k\|_{H^1}}_{:=\mathrm{Er}_{\mathrm{disc}}(u_N^k)} + \underbrace{\|u^k - u\|_{H^1}}_{:=\mathrm{Er}_{\mathrm{iter}}^k}, \tag{4.62}$$

where $u^k \in X$ is the approximate solution obtained after performing k iterations via the following iterative scheme

$$A_1 u^k + A_2 u^{k-1} = f. \tag{4.63}$$

In this definition, the term $u^k - u$ represents purely the iteration error without influence of finite-dimensional discretisation space.

The aim of our optimal path problem is, actually, to explore how to balance these two error sources such that the computation cost to achieve a given accuracy is optimized. It reveals from Proposition 4.2.1 that the error is bounded a posteriori with a measure of the residual. This leads naturally to the decomposition of residual such that each part represents one of the error sources. From the above GSR iterative scheme given by Equation (4.48), it is natural to define the discretisation residual as

$$R_{\mathrm{disc}}(u_N^k) \coloneqq A_1 u_N^k + A_2 u_N^{k-1} - f. \tag{4.64}$$

From Equation (4.59), it is clear that we have

$$\Pi_N R_{\mathrm{disc}}(u_N^k) = 0. \tag{4.65}$$

Then we define the iteration residual as the rest part in the total residual.

$$R_{\mathrm{iter}}(u_N^k) \coloneqq R(u_N^k) - R_{\mathrm{disc}}(u_N^k) = A_2(u_N^k - u_N^{k-1}). \tag{4.66}$$

After decomposing the residual into two part, we state the following lemma to show that, in some sense, the two residuals can represent respectively those two error sources.

Lemma 4.1.2. *For $N \in \mathbb{N}^*$, let the total residual be defined through* (4.36)*, let the discretisation residual be defined through* (4.64)*, let the iteration residual be defined through* (4.66)*, let $\Pi_N \colon X' \to X_N$ be the orthogonal projection operator onto space X_N and let $\Pi_N^{\perp} = \mathbf{I} - \Pi_N$ be its complementary.*
Let the iteration error and the discretisation error be defined through (4.61)*, then*

(1a) The total residual is bounded from below by the calculation error:

$$\gamma_a \|u_N^k - u\|_{H^1} \leqslant \|R(u_N^k)\|_{H^{-1}}, \tag{4.67}$$

where γ_a is the coercivity constant of $a \colon X \times X \to \mathbb{R}$ defined through (4.7)*.*

(1b) The iteration residual is bounded above and below by the iteration error:

$$\begin{aligned}\gamma_N \mathrm{Er}_{\mathrm{iter}}(u_N^k) \leqslant \|R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}} \leqslant &(\beta_N^2 + 2\|\Pi_N^\perp A_2\|_{H^1\to H^{-1}}^2)\mathrm{Er}_{\mathrm{iter}}(u_N^k) \\ &+ \|\Pi_N^\perp A_2\|_{H^1\to H^{-1}}\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1}),\end{aligned} \tag{4.68}$$

where γ_N and β_N are the coercivity and continuity constant of $\Pi_N A \Pi_N$.

(1c) The discretisation residual is bounded above by the discretisation error and the iteration error:

$$\begin{aligned}\|R_{\mathrm{disc}}(u_N^k)\|_{H^{-1}} \leqslant &\|\Pi_N^\perp A_1\|_{H^1\to H^{-1}}\mathrm{Er}_{\mathrm{iter}}(u_N^k) \\ &+ \|\Pi_N^\perp A_2\|_{H^1\to H^{-1}}\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1}) + \beta_a \mathrm{Er}_{\mathrm{disc}}^N,\end{aligned} \tag{4.69}$$

where β_a is the continuity constant of $a\colon X\times X\to\mathbb{R}$ defined through (4.6). Besides, when the iterations are enough, the discretisation residual is mainly bounded by the discretisation error.

Let the iteration error and the discretisation error be defined through (4.62), then

(2) The discretisation residual is bounded above by the discretisation error:

$$\|R_{\mathrm{disc}}(u_N^k)\|_{H^{-1}} \leqslant \|A_1\|_{H^1\to H^{-1}}\mathrm{Er}_{\mathrm{disc}}(u_N^k) + \|A_2\|_{H^1\to H^{-1}}\mathrm{Er}_{\mathrm{disc}}(u_N^{k-1}). \tag{4.70}$$

Proof. Picking $\widetilde{u_N} = u_N^k$ in (4.37) yields (4.67). For $N\in\mathbb{N}^*$, we have

$$\|\Pi_N R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}} \leqslant \|R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}} \leqslant \|\Pi_N R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}} + \|\Pi_N^\perp R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}}, \tag{4.71}$$

where in the above equality we make use of the fact that the L^2-orthogonal projection operator is also H^{-1}-orthogonal. It follows from Equations (4.65) and (4.66) that

$$\begin{aligned}\|\Pi_N R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}} &= \|\Pi_N(R(u_N^k) - R_{\mathrm{disc}}(u_N^k))\|_{H^{-1}} \\ &= \|\Pi_N R(u_N^k)\|_{H^{-1}} \\ &= \|\Pi_N A(u_N^k) - \Pi_N f\|_{H^{-1}} \\ &= \|\Pi_N A(u_N^k) - \Pi_N A(u_N)\|_{H^{-1}} \\ &= \|\Pi_N A\Pi_N(u_N^k - u_N)\|_{H^{-1}}\end{aligned}$$

where in the above derivation u_N is the discrete solution satisfying (4.60). Let γ_N and β_N be the coercivity and the continuity constant of $\Pi_N A\Pi_N$. We have

$$\gamma_N\|u_N^k - u_N\|_{H^1} \leqslant \|\Pi_N A\Pi_N(u_N^k - u_N)\|_{H^{-1}} \leqslant \beta_N\|u_N^k - u_N\|_{H^1}. \tag{4.72}$$

For the second right-hand side term of (4.71), from the second part of (4.66), we have

$$\begin{aligned}\|\Pi_N^\perp R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}} &= \|\Pi_N^\perp A_2(u_N^k - u_N^{k-1})\|_{H^{-1}} \\ &\leqslant \|\Pi_N^\perp A_2(u_N^k - u_N)\|_{H^{-1}} + \|\Pi_N^\perp A_2(u_N^{k-1} - u_N)\|_{H^{-1}} \\ &\leqslant \|\Pi_N^\perp A_2\|_{H^1\to H^{-1}}\|u_N^k - u_N\|_{H^1} + \|\Pi_N^\perp A_2\|_{H^1\to H^{-1}}\|u_N^{k-1} - u_N\|_{H^1}.\end{aligned} \tag{4.73}$$

Inserting Estimates (4.72) and (4.73) into (4.71) yields Estimate (4.68). For the discretisation

residual, thanks to (4.65) we have

$$\begin{aligned}
R_{\text{disc}}(u_N^k) =& A_1 u_N^k + A_2 u_N^{k-1} - f \\
=& \Pi_N^\perp A_1 u_N^k + \Pi_N^\perp A_2 u_N^{k-1} - \Pi_N^\perp f \\
=& \Pi_N^\perp A_1 u_N^k + \Pi_N^\perp A_2 u_N^{k-1} - \Pi_N^\perp A u_N + \Pi_N^\perp A u_N - \Pi_N^\perp f \\
=& \Pi_N^\perp A_1 (u_N^k - u_N) + \Pi_N^\perp A_2 (u_N^{k-1} - u_N) + \Pi_N^\perp (A u_N - f) \\
=& \Pi_N^\perp A_1 (u_N^k - u_N) + \Pi_N^\perp A_2 (u_N^{k-1} - u_N) + (A u_N - f),
\end{aligned}$$

where the last equality comes from (4.60). From Estimate (4.54), we have

$$\begin{aligned}
\|A u_N - f\|_{H^{-1}} &= \|A(u_N - u)\|_{H^{-1}} \\
&\leqslant \beta_a \|u_N - u\|_{H^1}.
\end{aligned}$$

Combining the above equations yields

$$\begin{aligned}
\|R_{\text{disc}}(u_N^k)\|_{H^{-1}} \leqslant& \|\Pi_N^\perp A_1\|_{H^1 \to H^{-1}} \text{Er}_{\text{iter}}(u_N^k) \\
&+ \|\Pi_N^\perp A_2\|_{H^1 \to H^{-1}} \text{Er}_{\text{iter}}(u_N^{k-1}) + \beta_a \text{Er}_{\text{disc}}^N.
\end{aligned} \tag{4.74}$$

Note that, from the above estimate, we know that when k goes to infinity, the iteration errors $\text{Er}_{\text{iter}}(u_N^{k-1})$ and $\text{Er}_{\text{iter}}(u_N^k)$ go to zero. Therefore, $\beta_a \text{Er}_{\text{disc}}^N$ is the dominant term in the above estimate when the number of iterations is large enough.

More simply, we have alternatively

$$\begin{aligned}
R_{\text{disc}}(u_N^k) =& A_1 u_N^k + A_2 u_N^{k-1} - f \\
=& A_1 u_N^k + A_2 u_N^{k-1} - (A_1 u^k + A_2 u^{k-1}) \\
=& A_1 (u_N^k - u^k) + A_2 (u_N^{k-1} - u^{k-1}),
\end{aligned}$$

with the iteration error and the discretisation error defined through (4.62), we deduce Estimate (4.70). □

4.1.4 Convergence and stability

In this part, we give a theoretical analysis about the convergence of the iterative scheme given by Equation (4.41). As mentioned before, the convergence of this iteration scheme is guaranteed by picking proper parameter ω such that the spectral radius of $\mathbf{P}_N$ is strictly smaller than 1: $\rho(\mathbf{P}_N) < 1$. Here we give the following convergence analysis result which is inspired by the work in [102].

Proposition 4.1.3. *Let the iteration matrix $\mathbf{P}_N$ be defined through Equation* (4.41)*, then there exist $\omega_1 < 1$ and $\omega_2 > 2$ such that for all $\omega_1 < \omega < \omega_2$ we have $\rho(\mathbf{P}_N) < 1$.*

Proof. Let us consider any eigenvalue of matrix $\mathbf{P}_N$:

$$\mathbf{P}_N \mathbf{v}^{\mathbf{P}} = \lambda^{\mathbf{P}} \mathbf{v}^{\mathbf{P}} \quad (\lambda^{\mathbf{P}} \in \mathbb{C}, \mathbf{v}^{\mathbf{P}} \in \mathbb{C}^{N+1}). \tag{4.75}$$

Recall the explicit expression of $\mathbf{P}_N$ in (4.40) and we have

$$\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N\right) \mathbf{v}^{\mathbf{P}} = -\lambda^{\mathbf{P}} (\mathbf{D}_N + \omega \mathbf{L}_N) \mathbf{v}^{\mathbf{P}}. \tag{4.76}$$

By multiplying on both sides with $\left(\mathbf{v}^{\mathbf{P}}\right)^T$, we get

$$\left(\mathbf{v}^{\mathbf{P}}\right)^T \left((1-\omega)\mathbf{L}_N + \mathbf{U}_N\right)\mathbf{v}^{\mathbf{P}} = -\lambda^{\mathbf{P}} \left(\mathbf{v}^{\mathbf{P}}\right)^T (\mathbf{D}_N + \omega\mathbf{L}_N)\mathbf{v}^{\mathbf{P}}. \tag{4.77}$$

We note incidentally that, in our case, $\mathbf{U}_N = \mathbf{L}_N^T$, so that we note by $\lambda_t = \left(\mathbf{v}^{\mathbf{P}}\right)^T \mathbf{L}_N\mathbf{v}^{\mathbf{P}} = \overline{\left(\mathbf{v}^{\mathbf{P}}\right)^T \mathbf{U}_N\mathbf{v}^{\mathbf{P}}}$. Besides, we introduce $\lambda_d = \left(\mathbf{v}^{\mathbf{P}}\right)^T \mathbf{D}_N\mathbf{v}^{\mathbf{P}}$. Hence, for any eigenvalue $\lambda^{\mathbf{P}}$ of $\mathbf{P}_N$ we have

$$\begin{aligned} \lambda^{\mathbf{P}} &= -\frac{(1-\omega)\lambda_t + \overline{\lambda_t}}{\lambda_d - \omega\lambda_t} \\ &= -\frac{(2-\omega)\,\mathrm{Re}(\lambda_t) + \omega\,\mathrm{Im}(\lambda_t)}{\lambda_d + \omega\,\mathrm{Re}(\lambda_t) + \omega\,\mathrm{Im}(\lambda_t)}. \end{aligned} \tag{4.78}$$

In addition, the ellipticity of the bilinear form a shows that $\lambda_d + 2\,\mathrm{Re}(\lambda_t) = \lambda_d + \lambda_t + \overline{\lambda_t} = \left(\mathbf{v}^{\mathbf{P}}\right)^T (\mathbf{D}_N + \mathbf{L}_N + \mathbf{U}_N)\,\mathbf{v}^{\mathbf{P}} = \left\langle Av^{\mathbf{P}}, v^{\mathbf{P}}\right\rangle_{X',X} > 0$ and from the expression of elements in matrix $\mathbf{D}_N$, we know that $\lambda_d > 0$. Hence, for $1 \leqslant \omega \leqslant 2$,

(i) if $\mathrm{Re}(\lambda_t) \geqslant 0$, then $\lambda_d + \omega\,\mathrm{Re}(\lambda_t) > \omega\,\mathrm{Re}(\lambda_t) \geqslant (2-\omega)\,\mathrm{Re}(\lambda_t) \geqslant 0$, hence, $|\lambda^{\mathbf{P}}| < 1$;

(ii) if $\mathrm{Re}(\lambda_t) < 0$, then $\lambda_d + \omega\,\mathrm{Re}(\lambda_t) > -(2-\omega)\,\mathrm{Re}(\lambda_t) > 0$, hence, $|\lambda^{\mathbf{P}}| < 1$.

In both cases the spectral radius of $\mathbf{P}_N$ is smaller than 1. By continuity, this remains true for $\omega_1 < \omega < \omega_2$ with some $\omega_1 < 1$ and $\omega_2 > 2$. □

Numerically, we plot the variation of $\rho(\mathbf{P}_N)$ as a function of ω for $N = 10$ in Figure 4.1. In addition, we test for $N = 10$: by setting $\omega = -0.2$ and $u_N^0 = 0$, we perform 30 GSR iterations and then plot the variation of associated norm $\|u_N^{k-1} - u_N^k\|_{H^1}$ for $k = 1, \ldots, 30$ in Figure 4.2 and of the energy $E(u_N^k)$ in Figure 4.3. From Figure 4.2, we note that at the beginning, the error decreases faster and then the decrease of the term $\|u_N^k - u_N^{k+1}\|_{H^1}$ is close to a straight line in log-scale. From the analysis in previous section, we know that this slope corresponds to the matrix norm $\|\mathbf{P}_N\|_{H^1 \to H^1}$ and the slope is a bit sharper at the beginning, which means that the iterations are more efficient at the beginning. This property will also be used when we propose the nearly optimal strategies in the following part.

4.2 Non linear case

Throughout this section, we assume the setting of Chapter 4.1 for the Sobolev space X and for function V and f. In this section, we consider the following energy minimization problem:

$$E^* = \min\{E(v) := \frac{1}{2}\int_0^{2\pi} (\nabla v)^2 + \frac{1}{2}\int_0^{2\pi} Vv^2 + \frac{1}{4}\int_0^{2\pi} v^4 - \langle f, v\rangle_{X',X}\,,\ v \in X\}. \tag{4.79}$$

And the corresponding variational problem that characterizes a function $u \in X$ which minimizes the above energy is defined as: Find $u \in X$ such that

$$\forall v \in X, \quad \int_0^{2\pi} \nabla u \cdot \nabla v + \int_0^{2\pi} Vuv + \int_0^{2\pi} u^3 v = \langle f, v\rangle_{X',X}\,. \tag{4.80}$$

or using the duality pairing notation

$$\forall v \in X, \quad \langle F(u), v\rangle_{X',X} = 0, \tag{4.81}$$

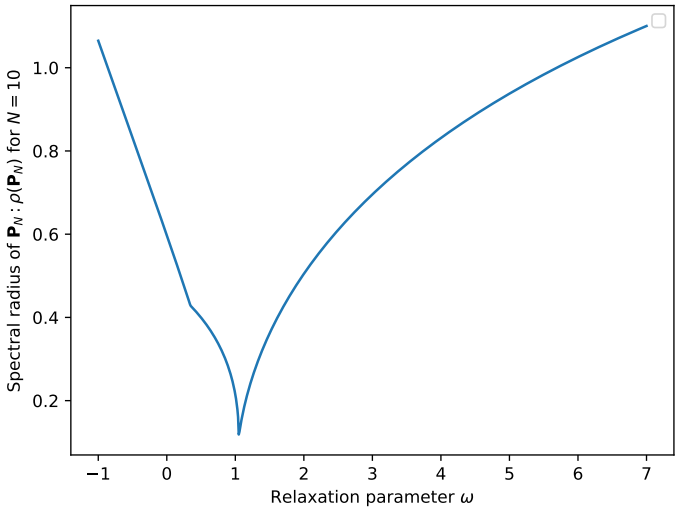


Figure 4.1: The variation of $\rho(\mathbf{P}_N)$ for $N = 10$ as a function of the relaxation parameter ω.

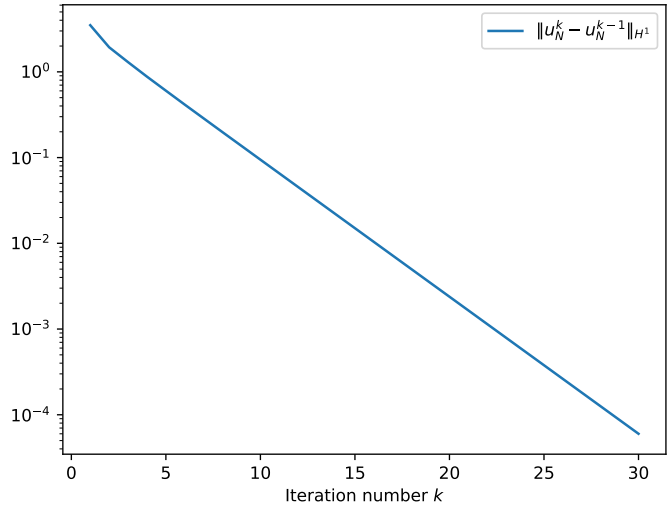


Figure 4.2: The variation of iteration increment $\|u_N^k - u_N^{k+1}\|_{H^1}$ for $N = 10$ as a function of iteration number k in log-scale.

where the nonlinear operator $F\colon X \to X'$ is defined as

$$\forall v \in X, \quad F(v) = -\Delta v + Vv + v^3 - f. \tag{4.82}$$

For proving the well-posedness of this problem, i.e., the existence and uniqueness of the

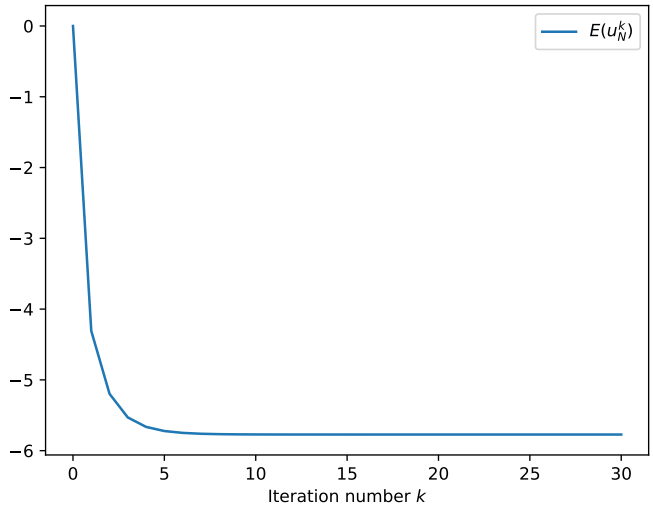


Figure 4.3: The variation of energy $E(u_N^k)$ for $N = 10$ as a function of iteration number k.

solution, we mainly make use of the property of nonlinear monotone operator. Therefore, we prove firstly that the operator $F\colon X \to X'$ is strongly monotone and that the energy functional defined in (4.79) is weakly coercive. Consequently, with a classical nonlinear functional analysis result, we prove the existence and uniqueness of the solution for (4.80), which, consequently, also minimizes the energy functional.

Proposition 4.2.1. *Let $E\colon X \to \mathbb{R}$ be defined through* (4.79)*, let $F\colon X \to X'$ be defined through* (4.82)*, which is the functional derivative of $E\colon X \to \mathbb{R}$. Then*

The minimization problem

$$\text{Find } u \in X \text{ such that } E(u) = \min_{v \in X} E(v), \tag{4.83}$$

and the operation equation

$$\text{Find } u \in X \text{ such that } F(u) = 0, \tag{4.84}$$

are equivalent. Besides, both of them have a unique solution.

In addition, let $u \in X$ be the solution of the above two problems. Then we have the following estimate:

$$\|u\|_{H^1} \leqslant \frac{1}{\gamma_a} \|f\|_{H^{-1}}, \tag{4.85}$$

where $\gamma_a = \min\{1, V_{\min}\}$ is the coercivity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.7)*.*

Proof. The proof is a direct application of Theorem 25.F in [104]. For showing the above well-posedness result, we need to prove two properties: $F\colon X \to X'$ is strongly monotone and $E\colon X \to \mathbb{R}$ is weakly coercive.

In the first step, we show the strong monotonicity of operator $F\colon X \to X'$, i.e., there exists constant $\gamma_F > 0$ such that

$$\forall w, v \in X, \quad \langle F(w) - F(v), w - v \rangle_{X',X} > \gamma_F \|w - v\|_{H^1}^2. \tag{4.86}$$

The proof of the above statement is straightforward: For any $w, v \in X$, we have

$$\begin{aligned}
\langle F(w) - F(v), w - v\rangle_{X',X} &= \int_0^{2\pi} \nabla(w-v)\cdot\nabla(w-v) + \int_0^{2\pi} V(w-v)^2 + \int_0^{2\pi} (w^3 - v^3)(w-v) \\
&= \int_0^{2\pi} (\nabla(w-v))^2 + \int_0^{2\pi} V(w-v)^2 + \int_0^{2\pi} (w^2 + wv + v^2)(w-v)^2 \\
&= \int_0^{2\pi} (\nabla(w-v))^2 + \int_0^{2\pi} V(w-v)^2 + \frac{1}{2}\int_0^{2\pi} \left(w^2 + v^2 + (w+v)^2\right)(w-v)^2 \\
&\geqslant \int_0^{2\pi} (\nabla(w-v))^2 + \int_0^{2\pi} V(w-v)^2 \\
&\geqslant \|\nabla(w-v)\|_{L^2}^2 + V_{\min}\|w-v\|_{L^2}^2 \\
&\geqslant \gamma_a \|w-v\|_{H^1}^2,
\end{aligned}$$

where $\gamma_a = \min\{1, V_{\min}\}$ is the coercivity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.7). This shows that $F\colon X \to X'$ is a strongly monotone operator.

In next step, we prove the weak coercivity of the energy functional $E\colon X \to \mathbb{R}$, i.e., $E(u) \to \infty$ as $\|u\|_{H^1} \to \infty$ on X.

For any $v \in X$, we have

$$\begin{aligned}
E(v) &= \frac{1}{2}\int_0^{2\pi} (\nabla v)^2 + \frac{1}{2}\int_0^{2\pi} Vv^2 + \frac{1}{4}\int_0^{2\pi} v^4 - \langle f, v\rangle_{X',X} \\
&\geqslant \frac{1}{2}\int_0^{2\pi} (\nabla v)^2 + \frac{1}{2}\int_0^{2\pi} Vv^2 - \langle f, v\rangle_{X',X} \\
&\geqslant \frac{1}{2}\gamma_a \|v\|_{H^1}^2 - \|f\|_{H^{-1}}\,\|v\|_{H^1} \\
&\geqslant \left[\frac{1}{2}\gamma_a \|v\|_{H^1} - \|f\|_{H^{-1}}\right]\,\|v\|_{H^1}.
\end{aligned}$$

From the above inequality, we deduce that when $\|v\|_{H^1} \to \infty$, the lower bound of $E(v)$ goes to infinity. This implies that $E(v)$ goes to infinity, which completes the proof of the weak coercivity.

By checking all the conditions of the above mentioned theorem, finally we obtain the well-posedness property of our nonlinear problem.

Picking $v = u$ and inserting it into the weak formulation (4.80) yields that

$$\begin{aligned}
\langle f, u\rangle_{X',X} &= \int_0^{2\pi} (\nabla u)^2 + \int_0^{2\pi} Vu^2 + \int_0^{2\pi} u^4 \\
&\geqslant \int_0^{2\pi} (\nabla u)^2 + V_{\min}\int_0^{2\pi} u^2 \\
&\geqslant \gamma_a \|u\|_{H^1}^2,
\end{aligned} \tag{4.87}$$

where $\gamma_a = \min\{1, V_{\min}\}$ is the coercivity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.7) and $V_{\min}$ is the lower bound of potential function V. From the above estimate, we deduce immediately

a rough estimation of $\|u\|_{H^1}$:

$$\begin{aligned} \|u\|_{H^1} &\leqslant \frac{1}{\gamma_a} \frac{\langle f, u\rangle_{X',X}}{\|u\|_{H^1}} \\ &\leqslant \frac{1}{\gamma_a} \|f\|_{H^{-1}}, \end{aligned} \tag{4.88}$$

which completes the proof. □

Remark 4.2.1.

Proposition 4.2.1 states the existence and uniqueness of the solution $u \in H^1(\mathbb{T})$ to the weak problem (4.81) *or the strong problem* (4.84)*. However, the solution u can be more regular.In fact, if $V \in H^r(\mathbb{T})$ $(r > \frac{1}{2})$ and $f \in H^s(\mathbb{T})$ $(s \geqslant -1)$, then $u \in H^\tau(\mathbb{T})$ with $\tau = 2 + \min\{r, s\}$ [10, 45, 63]. Specially, for our chosen potential function V and f defined in Remark 4.1.2, we have $u \in H^{1.55-\epsilon}(\mathbb{T})$ with $\epsilon > 0$. This property will be used later in the proof of the super convergence result.*

After proving the well-posedness of this problem, the next step is to resolve it numerically in X_N and get approximate solution u_N^k using iterative process. Similarly as in Equation (4.10), we write the unknown function as a sum of cosine functions and resolve for the truncated Fourier series coefficient in the discretisation space X_N.

Now, we introduce the discrete variational problem: For $N \in \mathbb{N}^*$, find $u_N \in X_N$ such that

$$\forall v_N \in X_N, \quad \int_0^{2\pi} \nabla u_N \cdot \nabla v_N + \int_0^{2\pi} V u_N v_N + \int_0^{2\pi} u_N^3 v_N = \langle f, v_N\rangle_{X',X}\,. \tag{4.89}$$

Using similar argument as in Proposition 4.2.1, the above questions has exactly one solution, which also minimizes the energy defined through (4.79) over the space X_N,

$$E(u_N) = \min_{v_N \in X_N} E(v_N). \tag{4.90}$$

4.2.1 *A priori* and *a posteriori* error estimation

The aim of this section is to derive the *a priori* and *a posteriori* analysis of the nonlinear problem. Compared to the linear case, we no longer have the continuity or the global Lipschitz continuity of our nonlinear operator $F\colon X \to X'$, therefore, some uncomputable constant like $\|u\|_{H^1}$ are needed in our *a priori* error estimator. Fortunately, we still have the strong monotonicity (c.f., the coercivity in the linear case) and it is straightforward to deduce our *a posteriori* error estimator. Now, as a first step, we give the *a priori* analysis about the convergence of the discrete solution $u_N \in X_N$ of problem (4.89) to the solution $u \in X$ of problem (4.80), which is based on the Céa's Lemma [19] and the proof of [38, Theorem 6.2.1].

Lemma 4.2.1. *Let $u \in X$ be the weak solution of problem* (4.80)*, for $N \in \mathbb{N}^*$, let $u_N \in X_N$ be the solution of discrete problem* (4.89)*, let γ_F be the monotonicity constant of operator $F\colon X \to X'$ and let $E\colon X \to \mathbb{R}$ be the energy functional defined though* (4.79)*. Then there exists constant C_1 such that*

$$\|u - u_N\|_{H^1} \leqslant C_1 \min_{v_N \in X_N} \|u - v_N\|_{H^1}. \tag{4.91}$$

If $u \in H^\tau(\mathbb{T})$ with $\tau > 1$, then there exists constant $C_2 > 0$ such that

$$\|u - u_N\|_{H^1} \leqslant C_2 N^{1-\tau}. \tag{4.92}$$

Moreover, there exists constant $C_3 > 0$ independent of N such that for any $r \leqslant \tau$, we have

$$\|u_N\|_{H^r} \leqslant C_3 \|u\|_{H^r}. \tag{4.93}$$

In addition, there exists constant C_4 such that

$$\frac{1}{2}\gamma_F \|u - u_N\|_{H^1}^2 \leqslant E(u_N) - E(u) \leqslant C_4 \|u - u_N\|_{H^1}^2. \tag{4.94}$$

Proof. Firstly, recalling the definition of the energy functional $E\colon X \to \mathbb{R}$, for any $v_N \in X_N$, we have

$$\begin{aligned} E(v_N) - E(u) =& \frac{1}{2}\int_0^{2\pi} (\nabla v_N)^2 + \frac{1}{2}\int_0^{2\pi} V v_N^2 + \frac{1}{4}\int_0^{2\pi} v_N^4 - \langle f, v_N\rangle_{X',X} \\ & - \frac{1}{2}\int_0^{2\pi} (\nabla u)^2 - \frac{1}{2}\int_0^{2\pi} V u^2 - \frac{1}{4}\int_0^{2\pi} u^4 + \langle f, u\rangle_{X',X} \\ =& \frac{1}{2}\int_0^{2\pi} (\nabla v_N)^2 + \frac{1}{2}\int_0^{2\pi} V v_N^2 + \frac{1}{4}\int_0^{2\pi} v_N^4 \\ & - \frac{1}{2}\int_0^{2\pi} (\nabla u)^2 - \frac{1}{2}\int_0^{2\pi} V u^2 - \frac{1}{4}\int_0^{2\pi} u^4 \\ & + \langle f, u - v_N\rangle_{X',X}. \end{aligned}$$

Here, with the property that $u \in X$ is the weak solution of problem (4.80), we have

$$\begin{aligned} E(v_N) - E(u) =& \frac{1}{2}\int_0^{2\pi} (\nabla v_N)^2 + \frac{1}{2}\int_0^{2\pi} V v_N^2 + \frac{1}{4}\int_0^{2\pi} v_N^4 \\ & - \frac{1}{2}\int_0^{2\pi} (\nabla u)^2 - \frac{1}{2}\int_0^{2\pi} V u^2 - \frac{1}{4}\int_0^{2\pi} u^4 \\ & + \int_0^{2\pi} \nabla u \cdot \nabla u - v_N + \int_0^{2\pi} V u(u - v_N) + \int_0^{2\pi} u^3(u - v_N) \\ =& \frac{1}{2}\int_0^{2\pi} (\nabla v_N - \nabla u)^2 + \frac{1}{2}\int_0^{2\pi} V(v_N - u)^2 + \frac{1}{4}\int_0^{2\pi} (v_N^4 - 4u^3 v_N + 3u^4) \\ =& \frac{1}{2}\int_0^{2\pi} (\nabla v_N - \nabla u)^2 + \frac{1}{2}\int_0^{2\pi} V(v_N - u)^2 \\ & + \frac{1}{4}\int_0^{2\pi} (v_N - u)^2 (v_N + u)^2 + \frac{1}{2}\int_0^{2\pi} u^2 (v_N - u)^2. \end{aligned} \tag{4.95}$$

By picking $v_N = u_N$ the solution of discrete problem (4.89), we give the first upper bound estimate of the energy difference in (4.94).

$$E(u_N) - E(u) \geqslant \frac{1}{2}\int_0^{2\pi} (\nabla u_N - \nabla u)^2 + \frac{1}{2}\int_0^{2\pi} V(u_N - u)^2 \geqslant \frac{1}{2}\gamma_F \|u_N - u\|_{H^1}^2. \tag{4.96}$$

According to the variational principle, together with the strong continuity of E and the density of $\bigcup_{N=1}^{\infty} X_N$ in $H^r(\mathbb{T})$, we have

$$\lim_{N\to\infty} E(u_N) - E(u) = \lim_{N\to\infty} \inf_{v \in X_N} E(v) - E(u) = 0, \tag{4.97}$$

Additionally, with (4.96) we conclude that

$$\lim_{N\to\infty} \|u - u_N\|_{H^1} = 0. \tag{4.98}$$

From the above result, we deduce that the sequence $(u_N)_{N\in\mathbb{N}^*}$ is bounded, i.e., there exists a constant $C_u > 0$, such that

$$\forall N \in \mathbb{N}^*, \quad \|u_N\|_{H^1} \leqslant C_u.$$

After obtaining the boundedness property of u_N, now we give the error estimate of the approximate solution u_N. Recalling the strong monotonicity of the operator $F\colon X \to X'$, we have

$$(I) := \langle F(u) - F(u_N), u - u_N\rangle_{X',X} \geqslant \gamma_F \|u - u_N\|^2_{H^1}.$$

Besides, for any $v_N \in X_N$, with the property that $u \in X$ is the weak solution of problem (4.81) and that $u_N \in X_N$ is the discrete weak solution of problem (4.89), we have

$$\begin{aligned}
(I) =& \langle F(u) - F(u_N), u\rangle_{X',X} \\
=& \langle F(u) - F(u_N), u - v_N\rangle_{X',X} \\
=& \int_0^{2\pi} \nabla(u - u_N)\cdot\nabla(u - v_N) + \int_0^{2\pi} V(u - u_N)(u - v_N) \\
&+ \int_0^{2\pi} (u^2 + uu_N + u_N^2)(u - u_N)(u - v_N) \\
\leqslant& \int_0^{2\pi} |\nabla(u - u_N)|\cdot|\nabla(u - v_N)| + \int_0^{2\pi} V|u - u_N|\cdot|u - v_N| \\
&+ \|u^2 + uu_N + u_N^2\|_{L^\infty} \int_0^{2\pi} |u - u_N|\cdot|u - v_N| \\
\leqslant& \int_0^{2\pi} |\nabla(u - u_N)|\cdot|\nabla(u - v_N)| + \int_0^{2\pi} V|u - u_N|\cdot|u - v_N| \\
&+ (\|u\|^2_{L^\infty} + \|u\|_{L^\infty}\|u_N\|_{L^\infty} + \|u_N\|^2_{L^\infty}) \int_0^{2\pi} |u - u_N|\cdot|u - v_N| \\
\leqslant& \int_0^{2\pi} |\nabla(u - u_N)|\cdot|\nabla(u - v_N)| + \int_0^{2\pi} V|u - u_N|\cdot|u - v_N| \\
&+ C^2_{\mathrm{GN}}(\|u\|^2_{H^1} + \|u\|_{H^1}\|u_N\|_{H^1} + \|u_N\|^2_{H^1}) \int_0^{2\pi} |u - u_N|\cdot|u - v_N| \\
\leqslant& (\beta_a + 3C^2_{\mathrm{GN}}C_u^2)\|u - u_N\|_{H^1}\|u - v_N\|_{H^1},
\end{aligned}$$

where β_a is the continuity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.6) and C_{GN} is the Gagliardo Nirenberg type inequality constant[2]. Combining the above two estimates about the

[2]This constant depends on the domain of definition $[0, 2\pi]$ and a detailed derivation of this constant can be found in [38, p.165]. In our case, we have

$$\forall v \in X, \quad \|v\|_{L^\infty} \leqslant \left(\frac{1}{4\pi} + \sqrt{\frac{1}{16\pi^2} + 1}\right)^{\frac{1}{2}} \|v\|_{H^1}.$$

term(I), we deduce that

$$\|u - u_N\|_{H^1} \leqslant \frac{\beta_a + 3C_{\mathrm{GN}}^2 C_u^2}{\gamma_F} \|u - v_N\|_{H^1}, \tag{4.99}$$

which gives the *a priori* estimate of the discrete solution (4.91) with $C_1 = \frac{\beta_a + 3C_{\mathrm{GN}}^2 C_u^2}{\gamma_F}$. Let us recall that $\Pi_N \colon X' \to X_N$ is the orthogonal projection operator onto space X_N. Then, inserting $v_N = \Pi_N u$ into the above estimate yields that

$$\begin{aligned} \|u - u_N\|_{H^1} &\leqslant C_1 \|u - \Pi_N u\|_{H^1} \\ &\leqslant C_1 N^{1-\tau} \|u\|_{H^\tau}, \end{aligned} \tag{4.100}$$

from which we deduce Estimate (4.94) with $C_2 = C_1 \|u\|_{H^\tau}$.

For any $v_N \in X_N$, it's easy to deduce that there exists $c > 0$ independent of N such that

$$\|v_N\|_{H^r} \leqslant c N^{r-s} \|v_N\|_{H^s}, \tag{4.101}$$

where $r \geqslant s$. For any $r \leqslant \tau$, combining Estimate (4.101) with the *a priori* error estimate given by Equation (4.91) yields that

$$\begin{aligned} \|\Pi_N u - u_N\|_{H^r} \leqslant & c N^{r-1} \|\Pi_N u - u_N\|_{H^1} \\ \leqslant & c N^{r-1} (\|\Pi_N u - u\|_{H^1} + \|u - u_N\|_{H^1}) \\ \leqslant & c N^{r-1} (C_1 + 1) \|\Pi_N u - u\|_{H^1} \\ \leqslant & c (C_1 + 1) \|u\|_{H^r}, \end{aligned}$$

where in the last step we use similar argument as that used in Estimate (4.100). From the above derivation we deduce Estimate (4.93).

The last part of the proof is to show the upper bound of the energy difference estimate, which is deduced from (4.95).

$$\begin{aligned} E(u_N) - E(u) &\leqslant \frac{1}{2}\beta_a \|u - u_N\|_{H^1}^2 + \frac{1}{4}\|u + u_N\|_{L^\infty}^2 \|u - u_N\|_{H^1}^2 + \frac{1}{2}\|u\|_{L^\infty}^2 \|u - u_N\|_{H^1}^2 \\ &\leqslant (\frac{1}{2}\beta_a + \frac{3}{2} C_{\mathrm{GN}}^2 C_u^2) \|u - u_N\|_{H^1}^2, \end{aligned}$$

which completes the proof. □

In addition to the above classical *a priori* error estimate given by Equation (4.16), we also have the following classical L^2 error estimate. Similar to the Lemma 4.1.1 and Theorem 4.1.1 in the linear problem, in the nonlinear case, we also have the following following classical L^2 error estimate and super convergence result.

Lemma 4.2.2. *Let $u \in X$ be the weak solution of problem* (4.80) *and for $N \in \mathbb{N}^*$, let $u_N \in X_N$ be the solution of discrete problem* (4.89)*. Then there exits constant $C_1 > 0$ such that*

$$\|u - u_N\|_{L^2} \leqslant C_1 N^{-1} \|u - u_N\|_{H^1}, \tag{4.102}$$

Proof. The proof is essentially identical to the proof of Lemma 4.1.1 with some obvious modifi-

cations. We define the bilinear form as follows:

$$\forall v, w \in X, \quad a'(v,w) \coloneqq \int_0^{2\pi} \nabla v \cdot \nabla w + \int_0^{2\pi} V' v w, \tag{4.103}$$

where $V' = V + u^2 + u_N^2 + u u_N$. Using now Equations (4.80) and (4.89), we deduce that the above bilinear form maintains the Galerkin orthogonality:

$$\begin{aligned}\forall v_N \in X_N, \quad a'(u-u_N, v_N) &= \int_0^{2\pi} \nabla(u-u_N)\cdot\nabla v_N + \int_0^{2\pi} V(u-u_N)v_N \\ &\quad + \int_0^{2\pi} (u^2 + u_N^2 + u u_N)(u-u_N)v_N \\ &= \int_0^{2\pi} \nabla(u-u_N)\cdot\nabla v_N + \int_0^{2\pi} V(u-u_N)v_N + \int_0^{2\pi} (u^3 - u_N^3)v_N \\ &= 0.\end{aligned}$$

Besides, $V' \in L^\infty(\mathbb{T})$ with $\|V'\|_{L^\infty} \leqslant \|V\|_{L^\infty} + \|u\|_{L^\infty}^2 + \|u_N\|_{L^\infty}^2 + \|u\|_{L^\infty}^2 \, \|u_N\|_{L^\infty}^2$ and V' is bounded from below by the same lower bound $V_{\min} > 0$ of function V. Therefore, the new bilinear form is also continuous and coercive. Then the rest of the proof is identical to that for Lemma 4.2.2, simply replacing $a\colon X \times X \to \mathbb{R}$ by $a'\colon X \times X \to \mathbb{R}$. □

Theorem 4.2.1. *Let $u \in X$ be the weak solution of problem (4.80), for $N \in \mathbb{N}^*$, let $u_N \in X_N$ be the solution of discrete problem (4.89), let $\Pi_N\colon X' \to X_N$ be the orthogonal projection operator onto space X_N and let V be the potential function in problem (4.80), let $V \in H^r(\mathbb{T})$ for some $r > \frac{1}{2}$ be the potential function in problem (5.2), let $f \in H^s(\mathbb{T})$ for some $s \geqslant -1$ and let us denote by $\tau = \min\{r, 2+s, 2\}$. Then for N large enough:*

- *There exits constant $C_2 > 0$ such that*

$$\|\Pi_N u - u_N\|_{H^1} \leqslant C_2 N^{-\frac{\tau}{2}-1} \|u - u_N\|_{H^1}. \tag{4.104}$$

- *There exits constant $C_3 > 0$ such that*

$$\|\Pi_N u - u_N\|_{L^2} \leqslant C_3 N^{-\tau-1} \|u - u_N\|_{H^1}. \tag{4.105}$$

Specifically, for our chosen V and f defined in Remark 4.1.2, we have

$$\|\Pi_N u - u_N\|_{H^1} \leqslant C_2 N^{-\frac{7}{4}+\epsilon} \|u - u_N\|_{H^1} \tag{4.106}$$

and

$$\|\Pi_N u - u_N\|_{L^2} \leqslant C_3 N^{-\frac{5}{2}+\epsilon} \|u - u_N\|_{H^1}, \tag{4.107}$$

where $\epsilon > 0$.

Proof. The proof is essentially identical to the proof of Theorem 4.1.1 with the bilinear form defined through (4.103).

One direct consequence of the above change is the regularity of the new potential term $V' = V + u^2 + u_N^2 + u u_N$: for $V \in H^r$ ($\frac{1}{2} < r$) and $f \in H^s(\mathbb{T})$ ($s \geqslant -1$), from Remark 4.2.1 we know that then $u \in H^t(\mathbb{T})$ with $t = 2 + \min\{r, s\} \geqslant 1$. In addition, from (4.93) we deduce that $u_N \in H^t(\mathbb{T})$ and $\|u_N\|_{H^t}$ is uniformly bounded by $\|u\|_{H^t}$. Therefore, we have $V' \in H^q(\mathbb{T})$ for $q = \min\{r, t\} = \min\{r, 2+s\}$.

Considering the above regularity change and combining the proof in Theorem 4.1.1 yields the super convergence result. □

After showing the *a priori* error estimate, now we give the *a posteriori* error estimate with the nonlinear residual at hand: for any $\widetilde{u_N} \in X_N$, we define the residual as

$$R(\widetilde{u_N}) \coloneqq -\Delta\widetilde{u_N} + V\widetilde{u_N} + \widetilde{u_N}^3 - f. \tag{4.108}$$

The proof of the following *a posteriori* error estimate is similar to that for the linear case while we make use of the strong monotonicity of operator F instead of the coercivity.

Lemma 4.2.3. *Let $u \in X$ be the weak solution of problem* (4.80)*, for $N \in \mathbb{N}^*$, let $\widetilde{u_N} \in X_N$ be a numerical approximation solution of discrete problem* (4.89)*, let γ_F be the monotonicity constant of operator $F\colon X \to X'$ and let $E\colon X \to \mathbb{R}$ be the energy functional defined through* (4.79)*. Then we have*

$$\|\widetilde{u_N} - u\|_{H^1} \leqslant \frac{1}{\gamma_F}\|R(\widetilde{u_N})\|_{H^{-1}}. \tag{4.109}$$

We also have the following energy estimate:

$$0 \leqslant E(\widetilde{u_N}) - E(u) \leqslant \frac{1}{\gamma_F^2}\left(\frac{1}{2}\beta_a + \frac{1}{4}C_{\mathrm{GN}}^2(\|\widetilde{u_N}\|_{H^1} + \frac{1}{\gamma_a}\|f\|_{H^{-1}})^2 + \frac{C_{\mathrm{GN}}^2}{2\gamma_a^2}\|f\|_{H^{-1}}^2\right)\|R(\widetilde{u_N})\|_{H^{-1}}^2. \tag{4.110}$$

or

$$\begin{aligned} 0 \leqslant E(\widetilde{u_N}) - E(u) \leqslant & \frac{1}{\gamma_F^2}(\frac{1}{2}\beta_a + \frac{3}{2}C_{\mathrm{GN}}^2\|\widetilde{u_N}\|_{H^1}^2)\|R(\widetilde{u_N})\|_{H^{-1}}^2 \\ & + \frac{2C_{\mathrm{GN}}^2}{\gamma_F^3}\|\widetilde{u_N}\|_{H^1}\,\|R(\widetilde{u_N})\|_{H^{-1}}^3 + \frac{3C_{\mathrm{GN}}^2}{4\gamma_F^4}\|R(\widetilde{u_N})\|_{H^{-1}}^4. \end{aligned} \tag{4.111}$$

Proof. Firstly, we recall the strong monotonicity of the operator $F\colon X \to X'$, we have

$$(II) := \langle F(u) - F(\widetilde{u_N}), u - \widetilde{u_N}\rangle \geqslant \gamma_F\|u - \widetilde{u_N}\|_{H^1}^2.$$

Besides, we have

$$\begin{aligned} (II) &= \langle F(u) - F(\widetilde{u_N}), u - \widetilde{u_N}\rangle \\ &= \langle -F(\widetilde{u_N}), u - \widetilde{u_N}\rangle \\ &\leqslant \| - \Delta\widetilde{u_N} + V\widetilde{u_N} + (\widetilde{u_N})^3 - f\|_{H^{-1}}\|u - \widetilde{u_N}\|_{H^1}, \end{aligned}$$

where in the above equation we use the property that $u \in X$ is the weak solution of problem (4.80).

By combining the above two estimates of the term (II), we deduce that

$$\|\widetilde{u_N} - u\|_{H^1} \leqslant \frac{1}{\gamma_F}\|R(\widetilde{u_N})\|_{H^{-1}}.$$

Now we come up to the estimate of the energy. The positivity of the term $E(\widetilde{u_N}) - E(u)$ is clear

thanks to the variational principle. Then, from (4.95), we have

$$\begin{aligned}
E(\widetilde{u_N}) - E(u) =& \frac{1}{2}\int_0^{2\pi} (\nabla\widetilde{u_N} - \nabla u)^2 + \frac{1}{2}\int_0^{2\pi} V(\widetilde{u_N} - u)^2 \\
&+ \frac{1}{4}\int_0^{2\pi} (\widetilde{u_N} - u)^2(\widetilde{u_N} + u)^2 + \frac{1}{2}\int_0^{2\pi} u^2(\widetilde{u_N} - u)^2 \\
\leqslant& \frac{1}{2}\beta_a\|\widetilde{u_N} - u\|_{H^1}^2 + \frac{1}{4}C_{\mathrm{GN}}^2\|\widetilde{u_N} + u\|_{H^1}^2\|\widetilde{u_N} - u\|_{H^1}^2 + \frac{1}{2}C_{\mathrm{GN}}^2\|u\|_{H^1}^2\|\widetilde{u_N} - u\|_{H^1}^2 \\
\leqslant& (\frac{1}{2}\beta_a + \frac{1}{4}C_{\mathrm{GN}}^2(\|\widetilde{u_N}\|_{H^1} + \|u\|_{H^1})^2 + \frac{1}{2}C_{\mathrm{GN}}^2\|u\|_{H^1}^2)\|\widetilde{u_N} - u\|_{H^1}^2.
\end{aligned} \tag{4.112}$$

Inserting Estimates (4.85) and (4.109) into (4.112) yields the a *posteriori* energy error estimate given by Equation (4.110). In addition, by taking the argument u a line back with above set of inequalities, we can go on and bound the term $\|u\|_{H^1}$ using the residual and $\|\widetilde{u_N}\|_{H^1}$:

$$\begin{aligned}
E(\widetilde{u_N}) - E(u) \leqslant& \frac{1}{2}\beta_a\|\widetilde{u_N} - u\|_{H^1}^2 + \frac{1}{4}C_{\mathrm{GN}}^2\|\widetilde{u_N} + u\|_{H^1}^2\|\widetilde{u_N} - u\|_{H^1}^2 + \frac{1}{2}C_{\mathrm{GN}}^2\|u\|_{H^1}^2\|\widetilde{u_N} - u\|_{H^1}^2 \\
\leqslant& \frac{1}{2}\beta_a\|\widetilde{u_N} - u\|_{H^1}^2 + \frac{1}{4}C_{\mathrm{GN}}^2(2\|\widetilde{u_N}\|_{H^1} + \|\widetilde{u_N} - u\|_{H^1})^2\|\widetilde{u_N} - u\|_{H^1}^2 \\
&+ \frac{1}{2}C_{\mathrm{GN}}^2(\|\widetilde{u_N} - u\|_{H^1} + \|\widetilde{u_N}\|_{H^1})^2\|\widetilde{u_N} - u\|_{H^1}^2 \\
\leqslant& (\frac{1}{2}\beta_a + \frac{3}{2}C_{\mathrm{GN}}^2\|\widetilde{u_N}\|_{H^1}^2)\|\widetilde{u_N} - u\|_{H^1}^2 + 2C_{\mathrm{GN}}^2\|\widetilde{u_N}\|_{H^1}\,\|\widetilde{u_N} - u\|_{H^1}^3 \\
&+ \frac{3}{4}C_{\mathrm{GN}}^2\|\widetilde{u_N} - u\|_{H^1}^4.
\end{aligned}$$

Inserting the a *posteriori* error estimate for the solution into the above expression yields (4.111), which completes the proof. □

4.2.2 Iteration scheme and analysis

As in the linear case, in this section, we will give details of the numerical resolution of this problem. Based on the discretized weak problem (4.89), where $u_N = \sum_{0\leqslant j<N}(\widehat{u_N})_j e_j$ and selecting the basis vectors $\{e_i\}_{0\leqslant i<N}$ as test functions v_N, we obtain a system of nonlinear equations represented in the following matrix form

$$\mathbf{A}_N(\mathbf{u}_N)\mathbf{u}_N = \mathbf{b}_N, \tag{4.113}$$

with $\mathbf{u}_N$ being the unknown vector of coefficients $\mathbf{u}_N = [(\widehat{u_N})_0, (\widehat{u_N})_1, \ldots, (\widehat{u_N})_N]^T$. Unlike the linear case, in the nonlinear problem, the matrix $\mathbf{A}_N$ is not constant and depends on $\mathbf{u}_N$ with a term generated by $\int_0^{2\pi} u_N^3 v_N$ that we denote by $\mathbf{S}_N(\mathbf{u}_N)$. For the rest part, which is the same as $\mathbf{A}_N$ in (4.39) for the linear case, we still keep the notation of $\mathbf{U}_N, \mathbf{D}_N$ and $\mathbf{L}_N$, i.e.,

$$\mathbf{A}_N(\mathbf{u}_N) = \mathbf{U}_N + \mathbf{D}_N + \mathbf{L}_N + \mathbf{S}_N(\mathbf{u}_N).$$

In the nonlinear case, we still use the Gauss-Seidel-Relaxation(GSR) iterative scheme. Starting from an initial guess $\mathbf{u}_N^0$, in each iteration step the new vector $\mathbf{u}_N^{k+1}$ is calculated from $\mathbf{u}_N^k$

via:

$$\begin{aligned}\mathbf{u}_N^{k+1} =&\xi\left(-(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N+\mathbf{U}_N+\mathbf{S}_N(\mathbf{u}_N^k)\right)\mathbf{u}_N^k+(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\mathbf{b}_N\right)\\&+(1-\xi)\mathbf{u}_N^k,\end{aligned}\tag{4.114}$$

with ω being the relaxation factor and $0<\xi<1$ being a damping factor. Similarly, by defining $\mathbf{P}_N(\mathbf{u}_N^k) := -(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N+\mathbf{U}_N+\mathbf{S}_N(\mathbf{u}_N^{k-1})\right)$ and $\mathbf{q}_N := (\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\mathbf{b}_N$, the above expression can be simplified as

$$\mathbf{u}_N^{k+1}=\xi\left(\mathbf{P}_N(\mathbf{u}_N^k)\mathbf{u}_N^k+\mathbf{q}_N\right)+(1-\xi)\mathbf{u}_N^k.$$

Remark 4.2.2 (Addition of the damping term)**.**

- *For the nonlinear problem, if we just make use of the GSR scheme like the one used for the linear problem, it will be written as follows:*

$$\mathbf{u}_N^{k+1}=-(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N+\mathbf{U}_N+\mathbf{S}_N(\mathbf{u}_N^k))\mathbf{u}_N^k+(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\mathbf{b}_N.\tag{4.115}$$

  However, the above scheme may not converge even by picking proper value of relaxation factor ω, i.e., we select a 'good' ω such that the spectral radius of $\mathbf{P}_N(\mathbf{u}_N^k)$ is between 0 and 1 for each iteration, the odd subsequence $(u_N^{2j+1})_{j\in\mathbb{N}}$ and even subsequence $(u_N^{2j})_{j\in\mathbb{N}}$ generally converge respectively to u_{odd} and u_{even}, but $u_{odd}\neq u_{even}$. This phenomenon also occurs when we apply the Roothaan algorithm to resolve the Hartree-Fock equations [17]. This is why we add an additional damping factor in the iteration process and in this way, the convergence of this algorithm is assured.

- *Here we don't split matrix $\mathbf{A}_N(\mathbf{u}_N)$ directly into the diagonal, upper and lower triangular parts. This is feasible but for the simplicity of the following residual analysis, we separate the nonlinear term $\mathbf{S}_N(\mathbf{u}_N)$ from the matrix $\mathbf{A}_N(\mathbf{u}_N)$ and divide the rest matrix into three parts: $\mathbf{D}_N$, $\mathbf{U}_N$ and $\mathbf{L}_N$.*

Compared to the linear case, in the nonlinear iteration scheme, there is the additional nonlinear term $\mathbf{S}_N(\mathbf{u}_N^k)$ in matrix $\mathbf{P}_N(\mathbf{u}_N^k)$. Hence, there is a slight difference in the convergence analysis. Compared to relation (4.42) and (4.43) in the linear case, the converged solution $\mathbf{u}_N^\infty$ of the nonlinear problem satisfies

$$\begin{aligned}\mathbf{u}_N^{\infty} =&\xi\left(-(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N+\mathbf{U}_N+\mathbf{S}_N(\mathbf{u}_N^{\infty})\right)\mathbf{u}_N^{\infty}+(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\mathbf{b}_N\right)\\&+(1-\xi)\mathbf{u}_N^{\infty}.\end{aligned}\tag{4.116}$$

Subtracting Equation (4.114) by (4.116) yields that

$$\begin{aligned}\mathbf{u}_N^{k+1}-\mathbf{u}_N^{\infty} =&\xi\left(-(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N+\mathbf{U}_N\right)(\mathbf{u}_N^k-\mathbf{u}_N^{\infty})\right.\\&\left.-(\mathbf{D}_N+\omega\mathbf{L}_N)^{-1}\left(\mathbf{S}_N(\mathbf{u}_N^k)\mathbf{u}_N^k-\mathbf{S}_N(\mathbf{u}_N^{\infty})\mathbf{u}_N^{\infty}\right)\right)+(1-\xi)(\mathbf{u}_N^k-\mathbf{u}_N^{\infty}).\end{aligned}\tag{4.117}$$

Noting that the matrix multiplication $\mathbf{S}_N(\mathbf{u}_N^k)\mathbf{u}_N^k$ in fact produces the vector containing the Fourier coefficients of function $\Pi_N\,\pi\cdot(u_N^k)^3$ and $\mathbf{S}_N(\mathbf{u}_N^\infty)\mathbf{u}_N^\infty$ corresponds to function $\Pi_N\,\pi\cdot(u_N^\infty)^3$. Therefore, we have the correspondence between the term $\mathbf{S}_N(\mathbf{u}_N^k)\mathbf{u}_N^k-\mathbf{S}_N(\mathbf{u}_N^\infty)\mathbf{u}_N^\infty$ and function $\Pi_N\ \pi\cdot\left((u_N^k)^3-\Pi_N\pi\cdot(u_N^\infty)^3\right)\ =\ \Pi_N\ \pi\cdot\left((u_N^k)^2+u_N^ku_N^\infty+(u_N^\infty)^2\right)\left(u_N^k-u_N^\infty\right)$. By denoting

$\mathbf{S}'_N(u_N^k, u_N^\infty)$ the matrix form of the function $(u_N^k)^2 + u_N^k u_N^\infty + (u_N^\infty)^2$, we deduce that

$$\begin{aligned}\mathbf{u}_N^{k+1} - \mathbf{u}_N^\infty =& \xi \left(-(\mathbf{D}_N + \omega \mathbf{L}_N)^{-1} \left((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}'_N(u_N^k, u_N^\infty)\right)(\mathbf{u}_N^k - \mathbf{u}_N^\infty)\right) \\ &+ (1-\xi)(\mathbf{u}_N^k - \mathbf{u}_N^\infty).\end{aligned} \tag{4.118}$$

By denoting $\mathbf{P}'_N(\mathbf{u}_N^k, \mathbf{u}_N^\infty) \coloneqq -(\mathbf{D}_N + \omega \mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}'_N(u_N^\infty, u_N^\infty)\right)$, we obtain

$$\mathbf{u}_N^{k+1} - \mathbf{u}_N^\infty = (\xi \mathbf{P}'_N(\mathbf{u}_N^k, \mathbf{u}_N^\infty) + (1-\xi)\mathbf{I}_N)(\mathbf{u}_N^k - \mathbf{u}_N^\infty).$$

Therefore, in the nonlinear case, the convergence of the iterative scheme (4.114) depends on the numerical solution $\mathbf{u}_N^k$ and it can't be achieved by simply picking proper values of ξ and ω. In practical calculation, we verify numerically that the spectral radius of the matrix $\xi \mathbf{P}'_N(\mathbf{u}_N^k, \mathbf{u}_N^\infty) + (1-\xi)\mathbf{I}_N$ is strictly smaller than one to ensure the convergence. Additionally, compare with the linear case, the convergence rate of the iteration scheme (4.114) also depends on the damping factor.

After defining the iterative scheme and analyzing its convergence, in the remaining part of this section, we aim at decomposing the residual according to our iterative scheme and showing the relation between the residual and the error. Recalling that $\Pi_N \colon X' \to X_N$ is the orthogonal projection operator onto space X_N, we have the following equivalent statements for the discretized form.

$$u_N \in X_N \text{ is solution of the weak problem (4.89)} \quad \Longleftrightarrow \quad u_N \in X_N,\ \Pi_N F(u_N) = 0.$$

Recall that the total residual is defined as $R(u_N^k) \coloneqq F(u_N^k)$. According to the expression of the iterative scheme given by Equation (4.114), we define the discretisation residual as follows

$$\begin{aligned}R_{\text{disc}}(u_N^k) \coloneqq& \frac{1}{\xi} A_1 u_N^k + A_2 u_N^{k-1} + (u_N^{k-1})^3 - (\frac{1}{\xi} - 1) A_1 u_N^{k-1} - f \\ =& \frac{1}{\xi} A_1 (u_N^k - u_N^{k-1}) + R(u_N^{k-1}),\end{aligned} \tag{4.119}$$

such that compared to the linear case, we maintain the property that

$$\Pi_N R_{\text{disc}}(u_N^k) = \frac{1}{\xi} \Pi_N \left(A_1 u_N^k + \xi \left(A_2 u_N^{k-1} + (u_N^{k-1})^3 - f\right) - (1-\xi) A_1 u_N^{k-1}\right) = 0, \tag{4.120}$$

where A_1 and A_2 are defined in Corollary 4.1.1.1 and we denote by $A = A_1 + A_2$. Then we define the iteration residual as the rest part in the total residual

$$\begin{aligned}R_{\text{iter}}(u_N^k) &\coloneqq R(u_N^k) - R_{\text{disc}}(u_N^k) \\ &= A_1 u_N^k + A_2 u_N^k + (u_N^k)^3 - f - \frac{1}{\xi} A_1 (u_N^k - u_N^{k-1}) - A_1 u_N^{k-1} - A_2 u_N^{k-1} - (u_N^{k-1})^3 + f \\ &= (1 - \frac{1}{\xi}) A_1 (u_N^k - u_N^{k-1}) + A_2 (u_N^k - u_N^{k-1}) + (u_N^k)^3 - (u_N^{k-1})^3.\end{aligned} \tag{4.121}$$

Now we state the lemma showing relation between error and residual.

Lemma 4.2.4. *For $N \in \mathbb{N}^*$, let the total residual be defined through* (4.108)*, let the discretisation residual be defined through* (4.119) *and let the iteration residual be defined through* (4.121)*. Let the iteration error and the discretisation error be defined through* (4.61)*, then*

(1a) The total residual is an upper bound for the calculation error:

$$\|u_N^k - u\|_{H^1} \leqslant \frac{1}{\gamma_F}\|R(u_N^k)\|_{H^{-1}}, \tag{4.122}$$

where γ_F is the monotonicity constant of operator $F\colon X \to X'$.

(1b) The iteration residual is bounded above and below by the iteration error:

$$\begin{aligned}\gamma_N \mathrm{Er}_{\mathrm{iter}}(u_N^k) \leqslant \|R_{\mathrm{iter}}(u_N^k)\|_{H^{-1}} \leqslant &\big(\beta_a + \frac{1}{\xi}\|A_1\|_{H^1\to H^{-1}} + C_{GN}^2(\|u_N^k\|_{L^2}^2 + \|u_N^k\|_{L^2}\|u_N^{k-1}\|_{L^2} \\ &+ \|u_N^{k-1}\|_{L^2}^2)\big)\left(\mathrm{Er}_{\mathrm{iter}}(u_N^k) + \mathrm{Er}_{\mathrm{iter}}(u_N^{k-1})\right),\end{aligned} \tag{4.123}$$

where γ_N is the coercivity, β_a is the continuity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.6).

(1c) The discretisation residual is bounded above by the discretisation error and the iteration error:

$$\begin{aligned}\|R_{\mathrm{disc}}(u_N^k)\|_{H^{-1}} \leqslant &\frac{1}{\xi}\|A_1\|_{H^1\to H^{-1}}\left(\mathrm{Er}_{\mathrm{iter}}(u_N^k) + \mathrm{Er}_{\mathrm{iter}}(u_N^{k-1})\right) + \beta_a\left(\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1}) + \mathrm{Er}_{\mathrm{disc}}\right) \\ &+ C_{GN}^2\left(\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1}) + \mathrm{Er}_{\mathrm{disc}}\right)\left(3\|u_N^{k-1}\|_{L^2}^2\right. \\ &+3\|u_N^{k-1}\|_{L^2}\left(\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1}) + \mathrm{Er}_{\mathrm{disc}}\right) \\ &\left.+\left(\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1}) + \mathrm{Er}_{\mathrm{disc}}\right)^2\right),\end{aligned} \tag{4.124}$$

where β_a is the continuity constant of $a\colon X \times X \to \mathbb{R}$ defined through (4.6)*. Besides, when the iterations are enough, the discretisation residual is mainly bounded by the discretisation error.*

Let the iteration error and the discretisation error be defined through (4.62)*, then*

(2) The discretisation residual is bounded above by the discretisation error:

$$\begin{aligned}\|R_{\mathrm{disc}}(u_N^k)\|_{H^{-1}} \leqslant &\frac{1}{\xi}\|A_1\|_{H^1\to H^{-1}}\left(\mathrm{Er}_{\mathrm{disc}}(u_N^k) + \mathrm{Er}_{\mathrm{disc}}(u_N^{k-1})\right) + \|A\|_{H^1\to H^{-1}}\mathrm{Er}_{\mathrm{disc}}(u_N^{k-1}) \\ &+ C_{GN}^2\,\mathrm{Er}_{\mathrm{disc}}(u_N^{k-1})\left(\mathrm{Er}_{\mathrm{disc}}(u_N^{k-1})^2 + \|u_N^{k-1}\|_{L^2}\mathrm{Er}_{\mathrm{disc}}(u_N^{k-1})^2 + \|u_N^{k-1}\|_{L^2}^2\right),\end{aligned} \tag{4.125}$$

where u^k could be obtained by setting an initial condition $u^0 \in X$ and calculated via the following iterative scheme:

$$\frac{1}{\xi}A_1(u^k - u^{k-1}) + Au^{k-1} + (u^{k-1})^3 = f. \tag{4.126}$$

Proof. Picking $\widetilde{u_N} = u_N^k$ in (4.109) yields (4.122). For the iteration residual, using Equation

(4.120), we deduce that

$$\begin{aligned}\|\Pi_N R_{\text{iter}}(u_N^k)\|_{H^{-1}} \leqslant \|R_{\text{iter}}(u_N^k)\|_{H^{-1}} \leqslant & \|\Pi_N R_{\text{iter}}(u_N^k)\|_{H^{-1}} + \|\Pi_N^\perp R_{\text{iter}}(u_N^k)\|_{H^{-1}} \\ \leqslant & \|\Pi_N \left(R(u_N^k) - R_{\text{disc}}(u_N^k)\right)\|_{H^{-1}} + \big\|(1-\frac{1}{\xi})\Pi_N^\perp A_1(u_N^k - u_N^{k-1}) \\ & + \Pi_N^\perp A_2(u_N^k - u_N^{k-1}) + \Pi_N^\perp \left((u_N^k)^3 - (u_N^{k-1})^3\right)\big\|_{H^{-1}},\end{aligned} \tag{4.127}$$

where $\Pi_N \colon X' \to X_N$ is the orthogonal projection operator onto space X_N and $\Pi_N^\perp = \mathbf{I} - \Pi_N$ is its complementary. It follows from Equation (4.120) that

$$\begin{aligned}\Pi_N \left(R(u_N^k) - R_{\text{disc}}(u_N^k)\right) &= \Pi_N R(u_N^k) \\ &= \Pi_N A u_N^k + \Pi_N (u_N^k)^3 - \Pi_N f \\ &= \Pi_N A u_N^k + \Pi_N (u_N^k)^3 - \Pi_N \left(A u_N + (u_N)^3\right) \quad (u_N \text{ is solution of (4.89)}) \\ &= \Pi_N A \Pi_N (u_N^k - u_N) + \Pi_N \left(\left((u_N^k)^2 + u_N^k u_N + u_N^2\right)(u_N^k - u_N)\right).\end{aligned}$$

It thus follows that

$$\begin{aligned}\|\Pi_N \left(R(u_N^k) - R_{\text{disc}}(u_N^k)\right)\|_{H^{-1}} \geqslant & \frac{\left\langle \Pi_N \left(R(u_N^k) - R_{\text{disc}}(u_N^k)\right), u_N^k - u_N \right\rangle_{X',X}}{\|u_N^k - u_N\|_{H^1}} \\ \geqslant & \frac{\left\langle \Pi_N A \Pi_N (u_N^k - u_N), u_N^k - u_N \right\rangle_{X',X}}{\|u_N^k - u_N\|_{H^1}} \\ & + \frac{\left\langle \Pi_N \left(\left((u_N^k)^2 + u_N^k u_N + u_N^2\right)(u_N^k - u_N)\right), u_N^k - u_N \right\rangle_{X',X}}{\|u_N^k - u_N\|_{H^1}}.\end{aligned}$$

Here we use the fact that the term $(u_N^k)^2 + u_N^k u_N + u_N^2$ is always positive to deduce that

$$\|\Pi_N R_{\text{iter}}(u_N^k)\|_{H^{-1}} \geqslant \gamma_N \|u_N^k - u_N\|_{H^1}, \tag{4.128}$$

where we recall that γ_N is the coercivity constant of $\Pi_N A \Pi_N$. For the continuity of $R_{\text{iter}}(u_N^k)$, we firstly rewrite

$$\begin{aligned}R_{\text{iter}}(u_N^k) =& (1-\frac{1}{\xi})A_1(u_N^k - u_N^{k-1}) + A_2(u_N^k - u_N^{k-1}) + (u_N^k)^3 - (u_N^{k-1})^3 \\ =& A(u_N^k - u_N^{k-1}) - \frac{1}{\xi}A_1(u_N^k - u_N^{k-1}) + \left((u_N^k)^2 + u_N^k u_N^{k-1} + (u_N^{k-1})^2\right)(u_N^k - u_N^{k-1}) \\ =& A(u_N^k - u_N) + A(u_N - u_N^{k-1}) - \frac{1}{\xi}A_1(u_N^k - u_N) - \frac{1}{\xi}A_1(u_N - u_N^{k-1}) \\ & \left((u_N^k)^2 + u_N^k u_N^{k-1} + (u_N^{k-1})^2\right)(u_N^k - u_N) + \left((u_N^k)^2 + u_N^k u_N^{k-1} + (u_N^{k-1})^2\right)(u_N - u_N^{k-1}).\end{aligned} \tag{4.129}$$

For any $v \in X$, we have

$$\begin{aligned}\left\langle (u_N^k)^2 (u_N^k - u_N), v \right\rangle = & \int_0^{2\pi} (u_N^k)^2 (u_N^k - u_N) v \\ \leqslant & \|v\|_{L^\infty} \|u_N^k - u_N\|_{L^\infty} \|u_N^k\|_{L^2}^2 \\ \leqslant & C_{\text{GN}}^2 \|v\|_{H^1} \|u_N^k\|_{L^2}^2 \text{Er}_{\text{iter}}(u_N^k),\end{aligned}$$

from which we deduce that

$$\|(u_N^k)^2(u_N^k-u_N)\|_{H^{-1}}=\sup_{v\in X,v\neq 0}\frac{\left\langle (u_N^k)^2(u_N^k-u_N),v\right\rangle}{\|v\|_{H^1}}\leqslant C_{\mathrm{GN}}^2\|u_N^k\|_{L^2}^2\mathrm{Er}_{\mathrm{iter}}(u_N^k). \tag{4.130}$$

By replacing $(u_N^k)^2$ with $(u_N^{k-1})^2$ or $u_N^k u_N^{k-1}$, we get similar result. Combing the above argument with Equation (4.129) yields (4.123).

For the discretisation residual. We have

$$\begin{aligned}
R_{\mathrm{disc}}(u_N^k)=&\frac{1}{\xi}A_1(u_N^k-u_N^{k-1})+R(u_N^{k-1})\\
=&\frac{1}{\xi}A_1(u_N^k-u_N^{k-1})+R(u_N^{k-1})-R(u)\\
=&\frac{1}{\xi}A_1(u_N^k-u_N^{k-1})+A(u_N^{k-1}-u)+(u_N^{k-1})^3-u^3\\
=&\frac{1}{\xi}A_1(u_N^k-u_N)+\frac{1}{\xi}A_1(u_N-u_N^{k-1})+A(u_N^{k-1}-u_N)+A(u_N-u)\\
&+\left((u_N^{k-1})^2+u_N^{k-1}u+u^2\right)(u_N^{k-1}-u)\\
=&\frac{1}{\xi}A_1(u_N^k-u_N)+\frac{1}{\xi}A_1(u_N-u_N^{k-1})+A(u_N^{k-1}-u_N)+A(u_N-u)\\
&+\left(3(u_N^{k-1})^2-3u_N^{k-1}(u_N^{k-1}-u)+(u_N^{k-1}-u)^2\right)(u_N^{k-1}-u).
\end{aligned} \tag{4.131}$$

Using similar argument as the one in Estimate (4.130), we obtain

$$\|(u_N^{k-1})^2(u_N^{k-1}-u)\|_{H^{-1}}\leqslant C_{\mathrm{GN}}^2\|u_N^{k-1}\|_{L^2}^2\left(\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1})+\mathrm{Er}_{\mathrm{disc}}\right), \tag{4.132}$$

$$\|(u_N^{k-1})(u_N^{k-1}-u)^2\|_{H^{-1}}\leqslant C_{\mathrm{GN}}^2\|u_N^{k-1}\|_{L^2}\left(\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1})+\mathrm{Er}_{\mathrm{disc}}\right)^2, \tag{4.133}$$

and

$$\|(u_N^{k-1}-u)^3\|_{H^{-1}}\leqslant C_{\mathrm{GN}}^2\|u_N^{k-1}\|_{L^2}^2\left(\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1})+\mathrm{Er}_{\mathrm{disc}}\right)^3. \tag{4.134}$$

Combining those arguments yields (4.124). When k goes to infinity, the iteration errors $\mathrm{Er}_{\mathrm{iter}}(u_N^{k-1})$ and $\mathrm{Er}_{\mathrm{iter}}(u_N^k)$ go to zero. Therefore the contribution of the discretisation residual is dominant in the above estimate when k is large enough.

More simply, we have alternatively

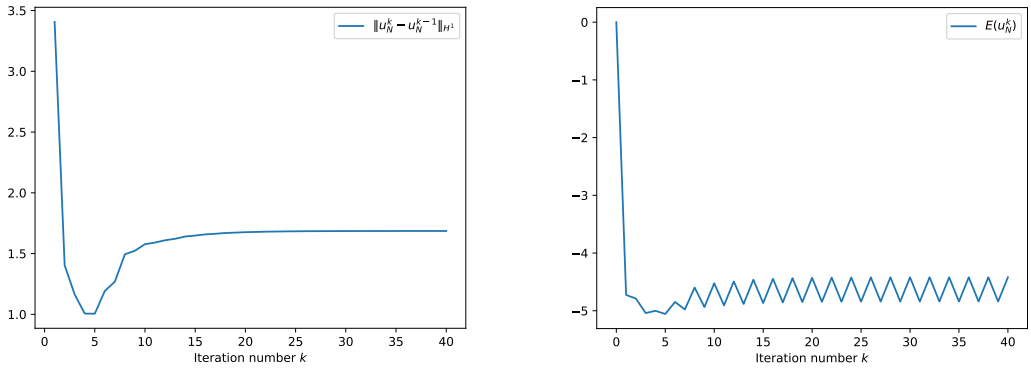
$$\begin{aligned}
R_{\mathrm{disc}}(u_N^k)=&\frac{1}{\xi}A_1(u_N^k-u_N^{k-1})+Au_N^{k-1}+(u_N^{k-1})^3-f\\
=&\frac{1}{\xi}A_1(u_N^k-u_N^{k-1})+Au_N^{k-1}+(u_N^{k-1})^3-\frac{1}{\xi}A_1(u^k-u^{k-1})-Au^{k-1}-(u^{k-1})^3\\
=&\frac{1}{\xi}A_1(u_N^k-u^k)-\frac{1}{\xi}A_1(u_N^{k-1}-u^{k-1})+A(u_N^{k-1}-u^{k-1})+(u_N^{k-1})^3-(u^{k-1})^3\\
=&\frac{1}{\xi}A_1(u_N^k-u^k)-\frac{1}{\xi}A_1(u_N^{k-1}-u^{k-1})+A(u_N^{k-1}-u^{k-1})\\
&+\left(3(u_N^{k-1})^2-3u_N^{k-1}(u_N^{k-1}-u^{k-1})+(u_N^{k-1}-u^{k-1})^2\right)(u_N^{k-1}-u^{k-1}).
\end{aligned}$$

Using similar argument as what we do in (4.132), (4.133) and (4.134) and combining with the new definition of discretisation error $\mathrm{Er}_{\mathrm{disc}}(u_N^k):=u_N^k-u^k$ yields Estimate (4.125), □

4.2.3 Convergence and stability

In this part, we aim at showing some numerical results related to the convergence of the GSR iterative scheme for our discrete nonlinear problem (4.89). As mentioned in remark 4.2.2, without the damping term the convergence of the GSR iterative scheme given by Equation (4.115) could not be guaranteed. The following numerical test reveals the oscillation of solution using (4.115):

For $N = 100$, given $u_N^0 = 0$ as the initial guess, we perform 40 GSR iterations according to Scheme (4.115) with $\omega = 1.1$ and plot the variation of error $\|u_N^k - u_N^{k-1}\|_{H^1}$ and energy $E(u_N^k)$ in Figure 4.4. We observe that the term $\|u_N^k - u_N^{k-1}\|_{H^1}$ doesn't converge to 0 and the oscillation of the energy is obvious from Figure 4.4B. However, we divide the numerical solution $(u_N^k)_{0\leqslant k\leqslant 40}$ into the odd part $(u_N^{2k-1})_{1\leqslant k\leqslant 20}$ and even part $(u_N^{2k})_{0\leqslant k\leqslant 20}$ and then plot respectively the variation of error and energy for these two parts in Figure 4.5 and Figure 4.6. We observe the convergence of the numerical solution respectively for the odd and even part. Besides, according to the variational principle, it is clear that none of these two subsequences converge to the discrete solution u_N of (4.89).

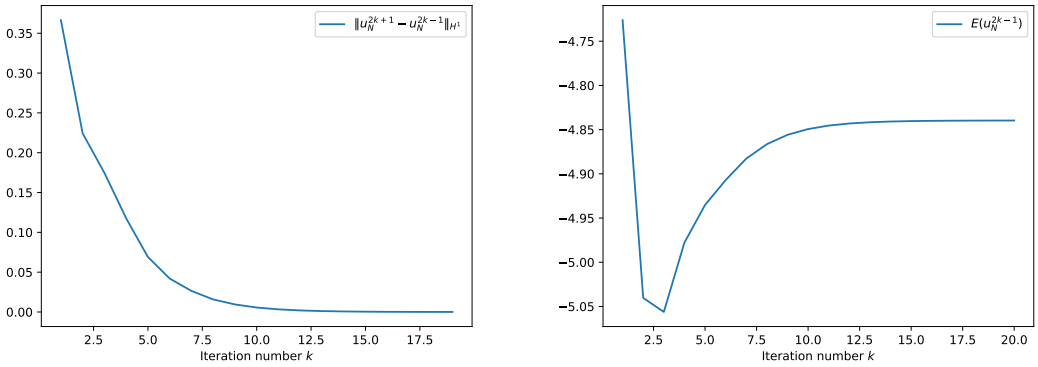


(A) The variation of $\|u_N^k - u_N^{k-1}\|_{H^1}$ for $N = 100$ as a function of iteration number k.

(B) The variation of $E(u_N^k)$ for $N = 100$ as a function of iteration number k.

Figure 4.4: The oscillation of nonlinear GSR scheme without damping term.

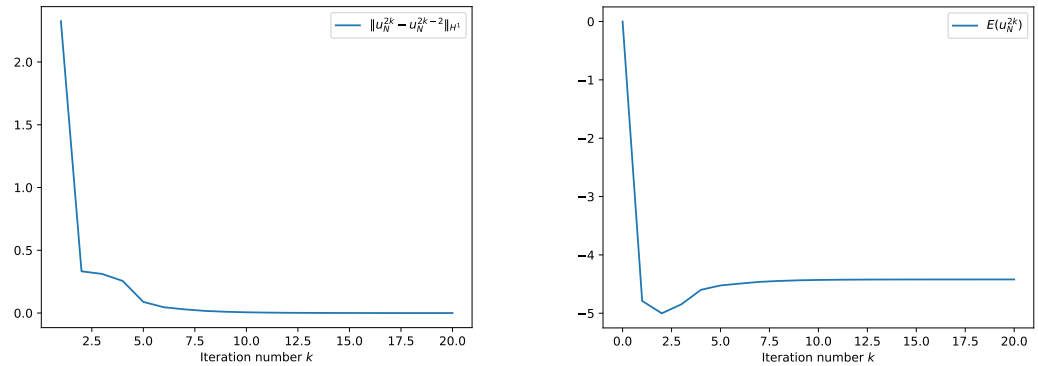
After explaining the necessity of adding a damping term into the GSR scheme given by Equation (4.114), now we show one numerical example using (4.114): For $N = 100$, given $u_N^0 = 0$ as the initial guess, we perform 40 GSR iterations with $\omega = 1.1$ and $\xi = 0.25$, then we plot the variation of error $\|u_N^k - u_N^{k-1}\|_{H^1}$ and energy $E(u_N^k)$ in Figure 4.7. After inserting a damping term into the GSR scheme, the convergence of the numerical solutions and of the energy are checked and the decrease rate of $\|u_N^k - u_N^{k-1}\|_{H^1}$ is also very close to a straight line in log-scale, which is the same as in the linear case. In the linear case, we know that the decrease rate of $\|u_N^k - u_N^{k-1}\|_{H^1}$ is close to the matrix norm $\|\mathbf{P}_N\|_{H^1\to H^1}$. But in the nonlinear case, matrix $\mathbf{P}_N(\mathbf{u}_N^k)$ depends on the current numerical solution. Only under the assumption that the numerical solution u_N^k is close enough to the discrete solution u_N, we can deduce that the decrease rate of the error $\|u_N^k - u_N^{k-1}\|_{H^1}$ is nearly a constant close to $\mathbf{P}_N(\mathbf{u}_N)$.



(A) The variation of $\|u_N^{2k+1} - u_N^{2k-1}\|_{H^1}$ for $N = 100$ as a function of iteration number k.

(B) The variation of $E(u_N^{2k-1})$ for $N = 100$ as a function of iteration number k.

Figure 4.5: The convergence of odd numerical solutions



(A) The variation of $\|u_N^{2k} - u_N^{2k-2}\|_{H^1}$ for $N = 100$ as a function of iteration number k.

(B) The variation of $E(u_N^{2k})$ for $N = 100$ as a function of iteration number k.

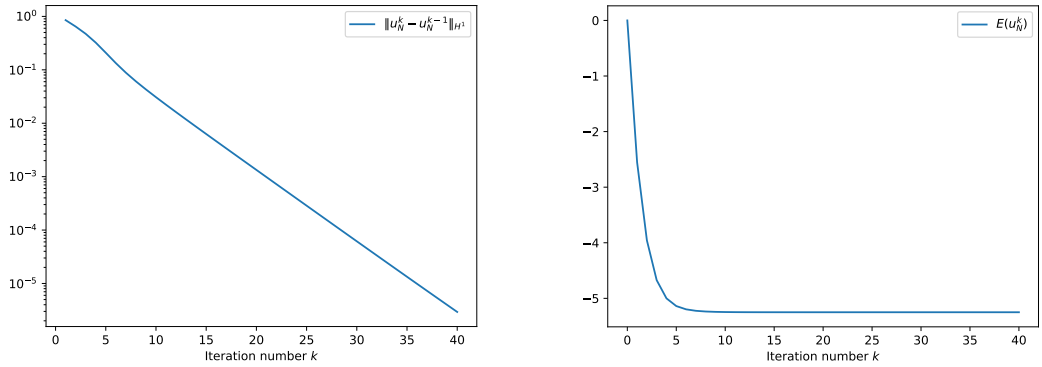
Figure 4.6: The convergence of even numerical solutions

4.3 Optimal path problem

4.3.1 Introduction

In the above two parts, we give the theoretical well-posedness analysis and some numerical convergence results of both the linear and the nonlinear problems. In this section, we explore the optimal error balance strategy such that these above two problems are numerically solved in an optimal way.

In most cases, the resolution of the above two problems involve firstly fixing a $N \in \mathbb{N}^*$, then constructing the discretisation space X_N and finally solving the problem over X_N. This discretisation number N which indicates the degree of freedom or the dimension of the discretisation space (the dimension is $N+1$) is chosen manually. If we pick $N = 10$, for example, the numerical



(A) The variation of $\|u_N^k - u_N^{k-1}\|_{H^1}$ for $N = 100$ as a function of iteration number k.

(B) The variation of $E(u_N^k)$ for $N = 100$ as a function of iteration number k.

Figure 4.7: The convergence of numerical solutions with damping term

solution of this problem is rather fast, however, the difference between our numerical solution and the true solution might be rather big. On the other hand, by picking a larger discretisation number N, e.g., $N = 100$, the numerical solution would be closer to the true solution, but it takes more computation resources to run the calculations. By only giving a target error bound, there is no evident clue for picking a proper N. It is possible to run a fast calculation but get a solution without satisfying the target bound or take much longer time for calculation and (luckily) get an accurate enough solution. Or even worse, we take a long-time calculation but still the final error is not satisfactory.

The aim of this work is to explore a new strategy to balance the computation resource and numerical results' accuracy. The idea is as follows: Firstly, we pick a small $N_{\min} \in \mathbb{N}^*$ as the starting point and initialize with $u^0_{N_{\min}} = 0$. Then we do several GSR iterations until we obtain a well converged solution in space $X_{N_{\min}}$. Next, we continue the numerical solution in a larger space $X_{N_1}(N_1 > N_{\min})$. Initializing the new series of iterations with the previous 'well converged solution' obtained in $X_{N_{\min}}$, we continue until obtaining, once again, well converged solution in space X_{N_1}. By repeating this process several times, finally we can get an accurate enough solution as desired. In addition, this may save computational resources provided that initializing iterations with previous well-converged solution diminishes the number of iterations needed to achieve the target accuracy.

After describing briefly the idea, the next step is to explore how to realize this idea. The first mission is to know how to switch between different discretisation spaces in a close-to-optimal way. Fortunately, the two problems studied are both periodic and we use cosine functions to approximate the unknown solution. Recall that in our numerical solution process, we obtain the coefficient vector $\mathbf{u}_N^k = [(\widehat{u_N^k})_0, (\widehat{u_N^k})_1, \ldots, (\widehat{u_N^k})_N]^T$ associated with the numerical approximation solution $u_N^k = \sum_{j=0}^N \widehat{u_N^k}_j e_j$. When we want to switch numerical solution from discretisation space X_N to space $X_{N'}(N' > N)$, we just add $N' - N$ zeros in the current coefficient vector $\mathbf{u}_N^k$ and define the resulting vector as $\mathbf{u}^0_{N'}$:

$$\mathbf{u}^0_{N'} = [(\widehat{u_N^k})_0, (\widehat{u_N^k})_1, \ldots, (\widehat{u_N^k})_N, 0, \ldots, 0]^T. \tag{4.135}$$

In this way, based on the solution obtained in X_N, we continue the calculation in a larger discretisation space $X_{N'}$ and at the same time the approximation solution isn't modified, i.e.,

$$u^0_{N'} = \sum_{j=0}^{N'} \widehat{(u^0_{N'})}_j e_j = \sum_{j=0}^{N} \widehat{(u^0_{N'})}_j e_j + \sum_{j=N+1}^{N'} \widehat{(u^0_{N'})}_j e_j = \sum_{j=0}^{N} \widehat{(u^k_N)}_j e_j = u^k_N.$$

Convention 4.3.1. *In the rest part of this section, we will frequently present the switch between two discretisation spaces. We will always denote by X_N the smaller discretisation space with dimension $N+1$ and by $X_{N'}$ the larger discretisation space with dimension $N'+1$. For simplicity, we will use discretisation number N to indicate corresponding discretisation space X_N, i.e., we will say the switch from N to N' to indicate the switch of numerical solution process from discretisation space X_N to discretisation space $X_{N'}$. Besides, we adapt the convention that $N' > N$ for the switch. And for $N \in \mathbb{N}^*$, we denote the number of iterations carried out in space X_N by k_N or s_N (which will be used in the Threshold Accepting simulation).*

After knowing how to switch between different discretisation spaces, the next question is to know how many iterations k_N should be performed in the fixed discretisation space X_N and how to choose next discretisation space $X_{N'}$. We name 'path' any possible calculation process that reduces the total error until the target tolerance is reached. A path collects the information about the choice of discretization spaces X_N and number of iterations k_N performed in corresponding discretization spaces and outputs it as an array. In addition, we aim at finding the **optimal path** minimizing the computational resources and achieving target accuracy. In our work, we set an error target for the approximate energy and explore the optimal way of finalizing the solution by achieving the target accuracy. There are a variety of possible paths and we use a probabilistic method, the threshold accepting method, to find the optimal path. In the rest of this section, we will introduce the TA method and present relevant mathematical settings. Finally we will show the optimal path result for both linear and nonlinear problems and discuss their differences. Here we remark that the way obtaining optimal path is not of practical use for solving the linear or the nonlinear source problem. The idea is to guide the intuition to define a possible near-optimal strategy.

4.3.2 Threshold Accepting method

Threshold Accepting method is proposed by Gunter and Tobias in [37], which is a variant of the well-known *Simulated Annealing*(SA) method [58], with its simpler structure and equal performance in some numerical examples. TA and SA are both probabilistic algorithms aiming at finding global optimum of a given function. Given an initial path, new neighboring path is generated as small random perturbation of the old path, then the 'quality' of those two paths are compared and the algorithm decides if the new path is kept for next iteration or not. The advantage of those two methods is that they accept new path having bad 'quality' with a tolerance, this is the key for finding the global optimum: sometimes the new path could be a local optimum and all its neighbors could be regarded as bad paths. By allowing the acceptance of bad paths, the algorithm can jump out from the local 'trap'. The difference between those two methods is the rule for acceptance of bad paths, which is known as *annealing schedule*. SA accepts bad path with a probability function which decreases with calculation time. However, TA accepts bad path with a deterministic tolerance which is replaced by a smaller one after a fixed number of iterations. Therefore, TA doesn't need to calculate probabilities to make the decision, which makes it easier to implement.

In our optimal path problem, we don't fully copy the algorithm in [37] but with some mod-

ifications. Here we state the general framework of TA method for our optimal path problem in Algorithm 1. After showing the algorithm, we will present in detail those functions appearing in the algorithm.

Algorithm 1 TA Algorithm for optimal path problem.

Input: Choose initial path S_{int} and initial threshold T, fix maximal number of iteration $K_{\max}$

while $T > 0$ **do**
  $k = 0$
  while $k < K_{\max}$ **do**
    $S_{\text{new}} = \text{Neighbor}(S)$
    if $\text{Cost}(S_{\text{new}}) < \text{Cost}(S) + T$ **then**
      $S \leftarrow S_{\text{new}}$
    end if
  end while
  Deacrease threshold T until $T = 0$
end while

Output: Optimal path S_{op}

As the first step, we define the error as the energy difference:

$$\varepsilon_N^k = E(u_N^k) - E(u_f). \tag{4.136}$$

Here, u_f is a converged very fine numerical solution, and in our problem, we set it as $u_f := u_{200}^{\infty}$. The notation ∞ doesn't means that we do infinite number of iterations, it is the converged numerical solution $u_N^{k^*}$ for fixed N and here $k^* = 200$ is sufficient. In our path, N varies between 3 and 100, according to the variational principle, we deduce that ε_N^k is always positive. Besides, we defined the discretisation energy error as follows.

$$\varepsilon_N^{\infty} = E(u_N^{\infty}) - E(u_f). \tag{4.137}$$

With the same variational argument, we know that ε_N^{∞} decreases as a function of N and in the general case it decreases, strongly. What's more, if we set the target error as, i.e., $\varepsilon = \varepsilon_N^{\infty}$, we can not arrive at the target accuracy without solving numerically the problem in discretization space $X_{N'}$ ($N' \geqslant N$). In the optimal path problem, we set the goal error as

$$\varepsilon_g = \frac{\varepsilon_{99}^{\infty} + \varepsilon_{100}^{\infty}}{2}. \tag{4.138}$$

We thus have $\varepsilon_g \geqslant \varepsilon_{100}^{\infty}$ and ε_g close to $\varepsilon_{100}^{\infty}$. Moreover, the total number of iterations for a possible path is limited such that the computational resources required for probabilistic calculation is limited. In addition, in this way, we make sure that we will need at least to perform one iteration for $N = 100$, which is the maximum discretisation number in the path. And any possible path has the following general form:

$$S = \{k_3, k_4, \ldots, k_{100}\}. \tag{4.139}$$

From the definition of goal error, we know that $k_{100} \geqslant 1$. In addition, the target accuracy is attained when we have $\varepsilon_{100}^k \leqslant \varepsilon_g$. Therefore, in addition to satisfy the accuracy condition, a path can be tested only when the number of iterations for $N = 100$ is minimized. This is equivalent to say that the iteration number k_{100} is a function of $k_3, k_4, \ldots, k_{99}$: for a random set $\{k_3, k_4, \ldots, k_{99}\}$, the corresponding k_{100} is uniquely defined as the smallest iteration number

for $N = 100$ such that the final error $\varepsilon_N^k \leqslant \varepsilon_g$. Therefore, the generation of a path is to firstly generate the integers set $\{k_3, k_4, \ldots, k_{99}\}$ and then add k_{100} which is obtained via calculation.

After defining properly the path, now we come to define the neighbor function which generates a new path from the old one with a small random perturbation. Firstly we pick a random number $3 \leqslant N \leqslant 99$, then for k_N, we generate the new neighbor path s'_N by picking a random integer in the interval $[\max\{k_N - 3, 0\}, k_N) \cup (k_N, \min\{k_N + 3, k_{\max}\}]$ such that we add a perturbation smaller than 3 iterations. Here, we set $k_{\max}$ as the maximal number of iterations, which is obtained by fixing the discretisation number as $N = 100$ and choosing $k_{\max}$ as the minimal number of iterations k such that $\varepsilon_{100}^k \leqslant \varepsilon_g$.

Another important function in the algorithm is the cost function that models the computation cost. It is based on the number of multiplications of two numbers. Therefore, in the linear problem, the cost of one iteration is $(N + 1)^2$ for discretisation number N (The dimension of X_N is $N + 1$) and the computational cost is proportional to the so-called cost function defined as follows:

$$\mathrm{Cost}(S) = \sum_{N=3}^{100} k_N (N + 1)^2. \tag{4.140}$$

For the nonlinear problem, in each iteration, the nonlinear cubic term needs to be updated and this part of cost is also counted. The cost function of the nonlinear problem is thus defined as:

$$\mathrm{Cost}(S) = \sum_{N=3}^{100} k_N (N + 1)^2 (N + 2). \tag{4.141}$$

The above Algorithm 1 is the sequential threshold acceptance method. In our problem, the dimension of search space is $(k_{\max} + 1)^{97}$, therefore it is slow to get the optimal path. To speed up this process, we run a parallel version of TA on 100 nodes and the algorithm is given in Algorithm 2. The threshold is initially set to be equal to the initial cost, and it is divided by 6 and after each loop the threshold is subtracted by $\frac{1}{6}$ of the initial cost.

Algorithm 2 Parallel TA Algorithm for optimal path problem.

Input: Choose initial path S_{int} and initial threshold T, fix maximal number of iteration $K_{\max}$ and number of nodes n in the network

while $T > 0$ **do**
  if the node is the central node **then**
    distribute initial path to $n - 1$ worker nodes
    run sequential TA algorithm with fixed threshold T
    get the simulation result and gather results from other worker nodes
    compare those results and choose the best one as initial path for next loop
  end if
  if the node is worker node **then**
    receive initial path from central node
    run sequential TA algorithm with fixed threshold T
    send the simulation result to the central node
  end if
  decrease threshold T until $T = 0$
end while

Output: Optimal path S

4.3.3 Optimal path result

In this subsection, we will present the optimal path result for both the linear case and the nonlinear case.

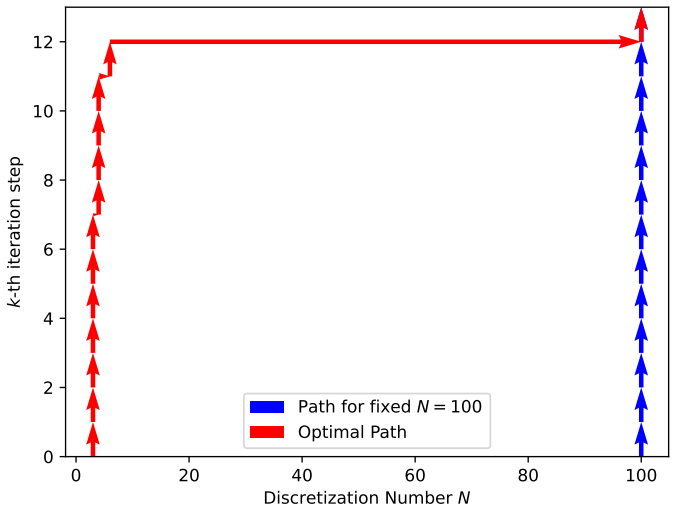


Figure 4.8: The optimal path for the linear case

Firstly we have the linear optimal path result in Figure 4.8. In this figure, we also add the path for fixed discretisation number $N = 100$ for comparison. For the optimal path and the fixed N path, we both need 13 iterations to achieve the target accuracy. The optimal path is $S_{\text{op}} = \{7, 4, 0, 1, 0, \ldots, 0, 1\}$: we do 7 iterations for $N = 3$, 4 iterations for $N = 4$, 1 iteration for $N = 6$ and finally 1 iteration for $N = 100$. The cost is $\text{Cost}(S_{\text{op}}) = 10,462$ compared with $\text{Cost}(S_{100}) = 132,613$ for fixed $N = 100$ path. From the expression of the cost function in (4.140), we know that the major contribution of the cost comes from the iteration for N in the range of 100, therefore, the solution path is optimized by decreasing as much as possible the number of iterations for large values of N while maintaining the required final accuracy. In this linear case, we get rather satisfactory result, this type of optimal path is what we can expect in the best case: to achieve the target accuracy, we need to do the same number of iterations,[3] but only one iterations for $N = 100$ is needed, most iterations can be realized with much smaller discretisation number N.

The result of nonlinear problem is shown in the Figure 4.9, and we also indicate the optimal path in Table 4.1. The cost of the optimal path is $\text{Cost}(S_{\text{op}}) = 10,216,786$. For the path with fixed $N = 100$, 15 iterations are needed to achieve the target accuracy and the cost is $\text{Cost}(S_{100}) = 15,607,530$, hence $\text{Cost}(S_{\text{op}})$ is about $\frac{2}{3}$ of $\text{Cost}(S_{100})$ and the optimal process does not help much. This is because nearly half of the iterations for $N = 100$ are still needed. Besides, unlike the linear optimal path, in the nonlinear case, iterations for small N doesn't play an important role in the optimal path, but iterations for a range of different N are required before jumping to $N = 100$.

[3]It seems that in most cases we can find manually some paths similar to the optimal one: 1)The number of iterations for this path is the same as for the$N = 100$ case. 2)All the iterations are done for N small except the last one iteration for $N = 100$. We are not sure if this is general.

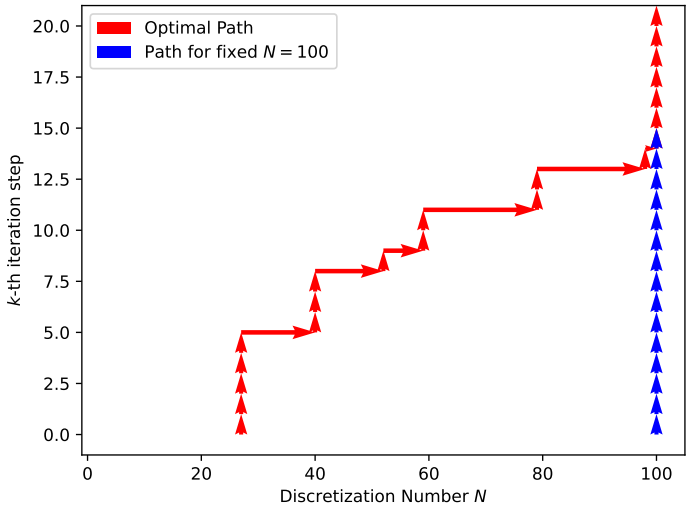


Figure 4.9: The optimal path for the nonlinear case with damping parameter ξ.

discretisation number N	27	40	52	59	79	98	100
Iteration number k_N	5	3	1	2	2	1	7

Table 4.1: The optimal path of the nonlinear case.

Remark 4.3.1 (Convergence of threshold accepting (TA) method)**.**

Theoretically, the probability that we obtain the global optimal result through TA method could approach 1 *under the condition that the number of nodes and iterations are large enough and that the decrease of threshold* T *is slow [1]. However, in practical calculation, the computation resources are limited and we can not guarantee that the output of TA calculation is truly the global optimal path for given problem. Or even worse, the output path could be manually checked that, unfortunately, it is not the optimal one.*

For example, an easy check can be implemented as follows: For any output path $S = \{k_3, k_4, \ldots, k_{100}\}$ *with at least one nonzero element* $k_N (4 \leqslant N \leqslant 99)$*, we pick one of those nonzero elements* $k_N (4 \leqslant N \leqslant 99)$*. Then we remove one iteration for* N*, i.e.,* $k'_N = k_N - 1$ *and add one iteration for* $N-1$*, i.e.,* $k'_{N-1} = k_{N-1} + 1$*. And we check if the new path* $S' = \{k_3, \ldots, k'_{N-1}, k'_N, \ldots, k_{100}\}$ *still satisfies the given accuracy or not. If the new path also works, then we obtain manually a more optimal path and it is evident that the previous path is not the optimal one.*

In our TA calculation, before decreasing the threshold T*, we will check if the output varies for bigger thresholds. The aim is to avoid being stuck in the local 'trap'. If the path* S *is always the same and we verify manually that it is not the optimal one. We will set* S' *obtained via the above described manual check as the new initial path and continue the TA calculation. Even finally the threshold arrives at* $T = 0$*, we still check manually if the output is the optimal one.*

For the nonlinear optimal path result in Figure 4.9, we observe that there is no iteration for small N*. In fact, the output TA calculation also contains nonzero elements* k_N *for small* N*. However, after checking manually the output path, we found out that even we remove all iterations for small* N*, the path still satisfies the target accuracy. Therefore, we conclude that*

small N doesn't play an important role in the nonlinear optimal path and plot the final manually-checked path in Figure 4.9.

From Figure 4.8, we observe that there is a jump from small N to $N = 100$ and then after 1 iteration the target accuracy is achieved. For small N, even we do enough iterations to get the nearly converged solution u_N, with the energy error being nearly equal to ε_N, there still exists an evident gap between ε_N and the goal error $\varepsilon_g = \frac{\varepsilon_{99}+\varepsilon_{100}}{2}$. However, 1 iteration for $N = 100$ is enough to cover this gap. This could be explained by the fact that the gap is not big enough or that the jump from small N to $N = 100$ is very efficient.

For checking if the 'good performance' of the linear optimal path result originates occasionally from a 'suitable' goal error, we add one more test for the linear optimal path problem: in this new test, we set the goal error as $\varepsilon_g = 0.01\varepsilon_{99} + 0.99\varepsilon_{100}$ and see if this time we can obtain similar path as the one in Figure 4.8.

Remark 4.3.2 (Optimized Path)**.**

*In the above additional test with $\varepsilon_g = 0.01\varepsilon_{99} + 0.99\varepsilon_{100}$, we don't run TA simulation to get optimal path result. Not only because TA simulation costs too much time, but also we declare again that the aim of this test is to see if we can reproduce a similar path as the one in Figure 4.8, but not to get the optimal path with new ε_g. We just enter manually the path $S = \{k_3, \ldots, k_{99}\}$ to see if we can achieve the target accuracy with $k_{100} = 1$. For example, we initialize the path with $k_N = k_{\max}$ with a small N (e.g. $N = 3$) and evaluate the number of iterations at $N = 100$ to achieve ε_g. By including gradually bigger N into the path, less iterations will be needed for $N = 100$ and we stop this process when $k_{100} = 1$. Next, we continue by replacing iterations for discretisation number N with smaller discretization numbers ($N - 1$, $N - 2$, etc) to get a manually cheaper path while maintaining $k_{100} = 1$ at the same time. Finally, we stop when we arrive at $N_{min=3}$ with optimized k_3 and we call the manually generated result as **optimized path** rather than optimal path.*

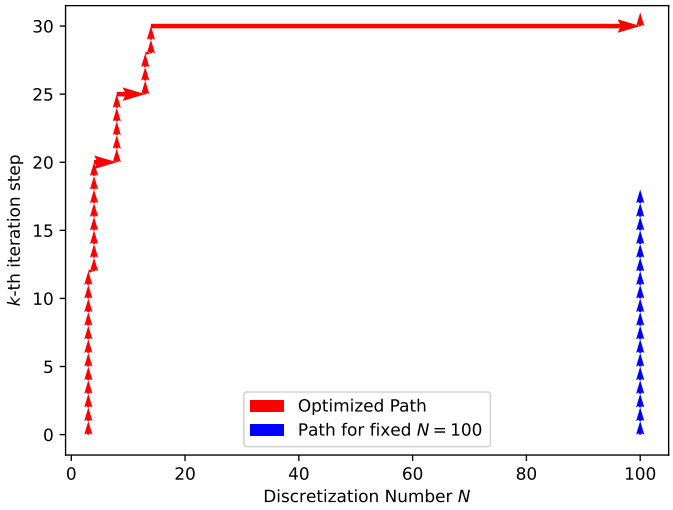


Figure 4.10: The optimized path for the linear case with $\varepsilon_g = 0.01\varepsilon_{99} + 0.99\varepsilon_{100}$

The optimized path is plotted in Figure 4.10. After setting new goal error to be closer to the discretisation error ε_{100}, we need 18 total number of iterations when we perform calculations with fixed $N = 100$. For the optimized path, we can still achieve the target error with only 1 iteration for $N = 100$, even though this time the range of small N is a bit larger, i.e., $3 \leqslant N \leqslant 14$ compared with $3 \leqslant N \leqslant 6$ in our original linear optimal path. Although, in the optimized path, more iterations are needed for small N, this part of additional computation costs are less important when compared to the costs of 1 iteration for $N = 100$. The result of this complementary test shows that the most important factor in the linear optimal is the jump from small N to $N = 100$: after getting a rather accurate simulation result for small N and jumping to $N = 100$, one iteration is enough to get the desired solution.

In this section, we show the optimal path result for both the linear and the nonlinear problems. In addition, we add one more test for the linear case by modifying the goal error. It seems that we could always realize an efficient jump in the linear case to reduce the calculation error more efficiently, which is impossible in the nonlinear case. At instance, we don't know much about the mechanism behind those two types of optimal paths. Still, we recall that the aim of this work is not only to know the optimal path result, but also to propose general strategies for any given accuracy requirement. Therefore, in next section, we will explore a bit about the mechanism generating two types of optimal paths and propose general strategies for both linear and nonlinear problems.

4.4 Complementary simulation results and analysis

4.4.1 Complementary simulation results

In previous section, we show the optimal paths for the linear and the nonlinear problem. In addition, we also add one more test to show the efficiency of the jump from small discretisation number N to $N = 100$ in the linear case. However, it seems that the jump between different discretisation numbers N (or more precisely, the switch of numerical solution in different discretisation spaces X_N) is less efficient in the nonlinear case. In this section, through a series of complementary tests, we explore the mechanism controlling the efficiency of jump between different discretisation numbers N and producing these two types of optimal paths.

Comparing the linear problem and the nonlinear problem, we find two differences in their iteration schemes: there are the nonlinear term $\mathbf{S}_N(\mathbf{u}_N)$ and the damping parameter $0 < \xi < 1$ in the nonlinear GSR scheme. According to this observation, we design the following tests and reveal the key factor by resuming these tests results.

Remark 4.4.1 (Nomenclature of complementary tests)**.**

The nomenclature principle of the following tests is as follows: Firstly, we clarify the type of problems to be solved: ***L*** *for the linear problem and* ***NL*** *for the nonlinear problem. Next, we distinguish the algorithm type used to solve the problem:* ***L*** *for the algorithm without damping term and* ***NL*** *for the algorithm with damping term. If there are more than one tests corresponding to the same type of problem and also the same type of algorithm, we will add letters as* ***A*** *or* ***B****. According to this principle, the linear optimal path result in Figure 4.8 corresponds to test for the type* ***L-L*** *and the nonlinear optimal path result in Figure 4.9 corresponds to test for the type* ***NL-NL****.*

Test **NL-L-A**, in this test, we resolve the nonlinear problem by transferring it into a linear problem.

$$-\Delta u + Vu + (u_{200}^{\infty})^2 u = f, \tag{4.142}$$

where u_{200}^{∞} is the converged solution for $N = 200$. The iteration scheme is the same as (4.40) and the matrix equation to be solved is

$$\mathbf{A}_N(u_{200}^{\infty})\mathbf{u}_N = \mathbf{b}_N. \tag{4.143}$$

In Test **NL-L-B**, we modify the matrix equation as

$$\mathbf{A}_N(\Pi_N u_{200}^{\infty})\mathbf{u}_N = \mathbf{b}_N, \tag{4.144}$$

where we recall that $\Pi_N : X \to X$ is the orthogonal projection operator onto space X_N.

The above two tests remove the influence of damping term. Test **L-NL** focus on problem without the existence of nonlinear term: in this test, for resolving the linear problem

$$-\Delta u + Vu = f, \tag{4.145}$$

we add damping term into the GSR iteration scheme

$$\begin{aligned}\mathbf{u}_N^{k+1} =& \xi\left(-(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N\right)\mathbf{u}_N^k + (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N\right) \\ &+ (1-\xi)\mathbf{u}_N^k,\end{aligned} \tag{4.146}$$

with $0 < \xi < 1$. As we stated in previous section, here our aim is to explore the mechanism controlling the efficiency of the jump but not to obtain the optimal path. Therefore, obtaining **optimized path** result is enough. These test results are plotted in Figure 4.11. We classify all these results into two types: **Type L** indicating the path type in figure 4.8 and **Type NL** indicating the path type in figure 4.9. From these test results, we observe that the existence of damping term plays the essential role for controlling the efficiency of the jump and thus producing different types of optimized paths. Furthermore, it seems that the linearity of the problem doesn't have important influence to the path types. In both the linear problem and the nonlinear problem, we can have two path types using different iterative schemes.

Now we know that the jump between different discretisation numbers N is more efficient for the algorithm without damping term. Therefore, we propose a new GSR algorithm for the nonlinear problem:

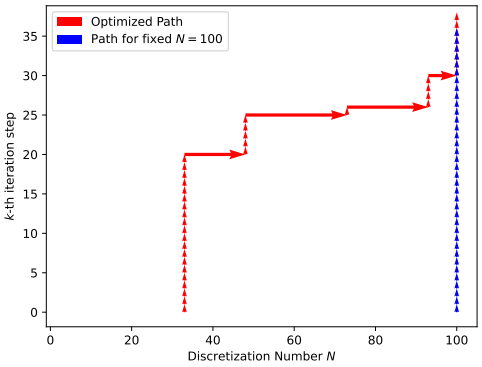
$$\mathbf{u}_N^{k+1} = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}_N\left(\frac{\mathbf{u}_N^{k-1} + \mathbf{u}_N^{k-2}}{2}\right)\right)\mathbf{u}_N^k + (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N. \tag{4.147}$$

After changing the iteration scheme, we can obtain converged result for the discrete nonlinear problem (4.89) and there is no oscillation of numerical solutions as shown in Figure 4.4. We name this case as Test **NL-L'** and we also plot its optimized path in Figure 4.11. We observe from the figure that after changing a new algorithm without damping term, we can produce optimized path of **Type L**, which is more efficient than the one of **Type NL**.

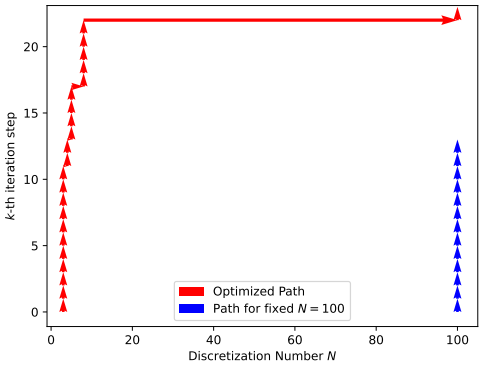
As a summary, we also list all different iteration schemes for different problems with corresponding path type in Table 4.2. In this section, we identify the key factor in the iterative scheme based on numerical calculations. In next section, we will give a theoretical analysis about the different behaviors of two path types.

4.4.2 Result analysis

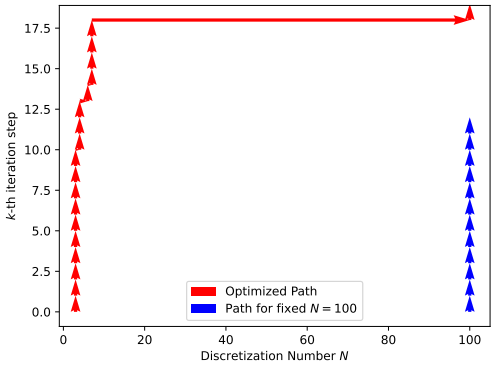
In the previous section, we offer some complementary tests and these test results show that the optimal path behavior relies strongly on the existence of the damping term. Thus, we explore



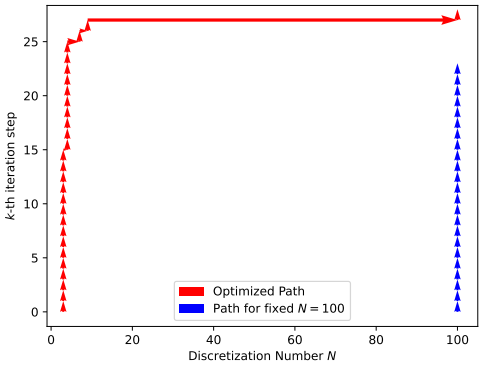
(A) Optimized path of test **L-NL**, we remark the same behavior for this problem as the one of **Type NL**, hence we deduce that the "damping parameter ξ is responsible of reducing the convergence speed.



(B) Optimized path of test **NL-L-A**, we remark here the same behavior for this linearized problem as the one of **Type L**, hence we deduce that the complexity of the equation is not responsible of reducing the convergence speed.



(C) Optimized path of test **NL-L-B**



(D) Optimized path of test **NL-L'**

Figure 4.11: The complementary tests results

firstly why the jump is so efficient for GSR iterative scheme without damping term.

Here, we recall the error analysis in section 4.1.3 and 4.2.2. In the first place, we focus on the linear problem, we know that the convergence rate of iteration algorithm is controlled by the matrix norm of matrix $\mathbf{P}_N$, which is a constant. However, from Figure 4.8 and Figure 4.10, it is clear that the first iteration after the jump reduces the iteration error efficiently. Therefore, here we do a simple test: We solve the linear problem using the iterative scheme given by Equation (4.40) without damping term. Firstly, we perform 10 iterations for $N = 3$ and we calculate the ratio $\frac{\|\mathbf{u}_N^{k+1} - \mathbf{u}_N^{\infty}\|_{H^1}}{\|\mathbf{u}_N^{k} - \mathbf{u}_N^{\infty}\|_{H^1}}$ for those 10 iterations. Those values vary between 0.5 and 0.69, which is reasonable because we have $0.698 < \|\mathbf{P}_3\|_{H^1 \to H^1} < 0.699$. Then we switch from $N = 3$ to $N' = 12$ and perform 4 iterations. We also calculate the ratio $\frac{\|\mathbf{u}_{N'}^{k+1} - \mathbf{u}_{N'}^{\infty}\|_{H^1}}{\|\mathbf{u}_{N'}^{k} - \mathbf{u}_{N'}^{\infty}\|_{H^1}}$ for those 4 iterations. The ratios are about $0.057, 0.451, 0.691, 0.693$ and we have $0.70 < \|\mathbf{P}_{12}\|_{H^1 \to H^1} < 0.701$. From

Table 4.2: List of different paths

Matrix Equation	Iteration scheme	Path type
$\mathbf{A}_N\mathbf{u}_N = \mathbf{b}_N$	$\mathbf{u}_N^{k+1} = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N + \mathbf{U}_N)\mathbf{u}_N^k + (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N$	**Type L**
$\mathbf{A}_N\mathbf{u}_N = \mathbf{b}_N$	$\mathbf{u}_N^{k+1} = -\xi(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N + \mathbf{U}_N)\mathbf{u}_N^k + \xi(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N + (1-\xi)\mathbf{u}_N^k$	**Type NL**
$\mathbf{A}_N(\Pi_N u_{200}^\infty)\mathbf{u}_N = \mathbf{b}_N$	$\mathbf{u}_N^{k+1} = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N + \mathbf{U}_N)\mathbf{u}_N^k + (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N$	**Type L**
$\mathbf{A}_N(u_{200}^\infty)\mathbf{u}_N = \mathbf{b}_N$	$\mathbf{u}_N^{k+1} = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N + \mathbf{U}_N)\mathbf{u}_N^k + (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N$	**Type L**
$\mathbf{A}_N(\mathbf{u}_N)\mathbf{u}_N = \mathbf{b}_N$	$\mathbf{u}_N^{k+1} = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}_N(\frac{\mathbf{u}_N^{k-1} + \mathbf{u}_N^{k-2}}{2}))\mathbf{u}_N^k + (\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N$	**Type L**
$\mathbf{A}_N(\mathbf{u}_N)\mathbf{u}_N = \mathbf{b}_N$	$\mathbf{u}_N^{k+1} = -\xi(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}_N(\mathbf{u}_N^{k-1}))\mathbf{u}_N^k + \xi(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\mathbf{b}_N + (1-\xi)\mathbf{u}_N^k$	**Type NL**

this simple test, we see that the ratio $\frac{\|\mathbf{u}_{N'}^{k+1} - \mathbf{u}_{N'}^\infty\|_{H^1}}{\|\mathbf{u}_{N'}^{k} - \mathbf{u}_{N'}^\infty\|_{H^1}}$ is very small for the first iteration after the jump, which means that this first iteration is very efficient and from the next iteration the ratio $\frac{\|\mathbf{u}_{N'}^{k+1} - \mathbf{u}_{N'}^\infty\|_{H^1}}{\|\mathbf{u}_{N'}^{k} - \mathbf{u}_{N'}^\infty\|_{H^1}}$ becomes closer and closer to $\|\mathbf{P}_{12}\|_{H^1\to H^1}$, which means that these iterations become less efficient.

After the jump from $N = 3$ to $N' = 12$, we have $\mathbf{u}_{12}^0 = \{(\widehat{\mathbf{u}_3^{10}})_0, \ldots, (\widehat{\mathbf{u}_3^{10}})_3, 0, \ldots, 0\}$. In general, after large enough number of iterations, we assume that the approximate solution $\mathbf{u}_3^k$ is close to the converged solution $\mathbf{u}_3^\infty$. In addition, providing that it is also close to $\Pi_3\mathbf{u}_{12}^\infty$ where Π_3 is the projection operator picking out the first four elements, the iteration error can thus be approximately expressed as $\mathbf{u}_{12}^\infty - \mathbf{u}_{12}^0 \approx \{0, \ldots, 0, (\widehat{\mathbf{u}_{12}^\infty})_4, \ldots, (\widehat{\mathbf{u}_{12}^\infty})_{12}\}$. Then for the first iteration after the jump, from Equation (4.43), we know that the matrix taking effect is not $\mathbf{P}_{12}$ but rather $\mathbf{P}_{12}(\mathbf{I} - \Pi_3)$. We have the norm $0.064 < \|\mathbf{P}_{12}(\mathbf{I} - \Pi_3)\|_{H^1\to H^1} < 0.065$, which explains why the first iteration after the jump is more efficient than next iterations for fixed N. In addition, we do one more test, we put $\Pi_3\mathbf{u}_{12}^\infty$ as input for $N = 3$, perform 1 iteration after the jump from $N = 3$ to $N' = 12$ and calculate the ratio $\frac{\|\mathbf{u}_{N'}^{1} - \mathbf{u}_{N'}^\infty\|_{H^1}}{\|\mathbf{u}_{N'}^{0} - \mathbf{u}_{N'}^\infty\|_{H^1}} = 0.0526$. This result verifies our explanation that the first iteration after a jump is more efficient because the matrix norm of the 'true' part $\|\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\|_{H^1\to H^1}$ is smaller than $\|\mathbf{P}_{N'}\|_{H^1\to H^1}$. Lastly, it follows from the super convergence result (4.26) in Theorem 4.1.1 that the term $\|u_N^\infty - \Pi_N u_{N'}^\infty\|_{H^1}$ decreases at a rate of $N^{-\frac{7}{4}+\epsilon}$ $(\epsilon > 0)$[4]. Therefore, this term is negligible as long as N and N' are large enough and in this linear source problem, we numerically checked that this assumption is satisfied.

In Table 4.2, there are four schemes with **Type L** path. For the linear case (type **L-L** or **NL-L**), we have three schemes with **Type L** path. For those three schemes, we verify that the above argument holds in each scheme. For the nonlinear case, the iterative scheme without the damping term also produces **Type L** optimized path. The convergence analysis is similar to the one in Chapter 4.2.2. By setting $\xi = 1$ and assuming that the algorithm is already well converged

[4]Here we checked numericaly that the decrease rate of this term is faster that N^{-2}.

before the jump, i.e., $u_N^k \approx u_N^{k+1}$, we obtain

$$\mathbf{u}_{N'}^{k+1} - \mathbf{u}_{N'}^{\infty} = \mathbf{P}'_{N'}(\mathbf{u}_N^k, \mathbf{u}_N^{\infty})(\mathbf{u}_{N'}^k - \mathbf{u}_{N'}^{\infty}),$$

where $\mathbf{P}'_{N'}$ is defined in (4.118). The jump efficiency is controlled by the term $\|\mathbf{P}'_{N'}(\mathbf{I} - \Pi_N)\|_{H^1 \to H^1}$. Similarly, in this case, we also checked numerically that the matrix norm $\|\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\|_{H^1 \to H^1}$ is smaller than $\|\mathbf{P}_{N'}\|_{H^1 \to H^1}$. The last remark is that for the nonlinear problem, we also have the super convergence result (4.106) in Theorem 4.2.1, the assumption that the term $\|u_N^{\infty} - \Pi_N u_{N'}^{\infty}\|_{H^1}$ is negligible is satisfied.

After explaining the efficiency of the jump in **Type L** path for both the linear and the nonlinear problems, now we take the linear problem **L-NL** as example to explain the behavior of **Type NL** path. In the case of **Type NL** path, we have a damping parameter ξ in the iteration scheme. In this situation, the jump isn't efficient anymore. Similarly to the analysis in Chapter 4.2.2, we firstly write the GSR iteration scheme as

$$\mathbf{u}_N^{k+1} - \mathbf{u}_N^{\infty} = (\xi \mathbf{P}_N + (1-\xi)\mathbf{I}_N)(\mathbf{u}_N^k - \mathbf{u}_N^{\infty}).$$

Therefore, by denoting that $\mathbf{P}''_N = \xi \mathbf{P}_N + (1-\xi)\mathbf{I}_N$, we obtain

$$\begin{aligned}
\|\mathbf{P}''_{N'}(\mathbf{I} - \Pi_N)\|^2_{H^1 \to H^1} =& \lambda_{\max}\left((\mathbf{T}_{N'}\mathbf{P}''_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{T}_{N'}\mathbf{P}''_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})\right) \\
=& \lambda_{\max}(\xi^2(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1}) \\
&+ (1-\xi)^2(\mathbf{T}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{T}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1}) \\
&+ \xi(1-\xi)(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{T}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1}) \\
&+ \xi(1-\xi)(\mathbf{T}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1}) \\
=& \lambda_{\max}(\xi^2(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1}) \\
&+ (1-\xi)^2(\mathbf{I} - \Pi_N) \\
&+ \xi(1-\xi)(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{I} - \Pi_N) \\
&+ \xi(1-\xi)(\mathbf{I} - \Pi_N)^T(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1}),
\end{aligned}$$

where the matrix $\mathbf{T}_{N'}$ is a diagonal matrix with the diagonal part $[(1+j^2)^{\frac{1}{2}}]_{0 \leqslant j \leqslant N'}$. In the above expression, as in the analysis of **Type L** path, we assume and have checked numerically for iterative scheme without damping term that the term $\|\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\|^2_{H^1 \to H^1} = \lambda_{\max}((\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1})^T(\mathbf{T}_{N'}\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\mathbf{T}_{N'}^{-1}))$ is small compared to $\|\mathbf{P}_{N'}\|^2_{H^1 \to H^1}$. Moreover, it is evident that the eigenvalue of the matrix $(\mathbf{I} - \Pi_N)$ are zero and one. In addition, for e.g., $\xi = 0.25$, combining those above arguments, we claim that the dominant term in the above expression is $\lambda_{\max}\left((1-\xi)^2(\mathbf{I} - \Pi_N)\right)$ and that $\|\mathbf{P}''_{N'}(\mathbf{I} - \Pi_N)\|_{H^1 \to H^1}$ is thus close to $1 - \xi = 0.75$. Here we check numerically the above statement: we calculate $0.757 < \|\mathbf{P}''_{12}(\mathbf{I} - \Pi_3)\|_{H^1 \to H^1} < 0.758$. This verifies our explanation: After adding a damping term, we need more iterations for $N = 100$ and 1 iteration is far from being enough. Besides, we need iterations for a range of different N to gradually decrease the discretisation error. Lastly, we also numerically check that the above explanation holds for the nonlinear iterative scheme with the damping term.

Here we give a brief summary to finalize this section: we have two optimal path types **Type L** and **Type NL**. The key factor determining the path behavior is the existence of the damping term. Without the damping term, i.e. the **Type L** case, the jump is efficient because the decrease of iteration error is not controlled, temporally, by the spectral radius of $\mathbf{P}_{N'}$ but $\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)$, which is smaller. For the **Type NL** case, the efficiency of the jump is degraded because of the existence of the damping term and the efficiency is mainly controlled by the parameter ξ.

4.5 Calculation strategy

In previous sections, we study the optimal path problem. Then by applying the TA method, we obtain optimal paths for the linear and the nonlinear problems. Next, we explore the mechanism of producing two kinds of optimal paths. By adding complementary tests and giving the detailed explanation, we reveal the critical impact of the damping term on the iterative scheme for determining the optimal path type. Now, in this section, we aim at proposing general strategies to solve efficiently these problems under a certain calculation goal.

From these optimal path results in Chapter 4.3 and the analysis in Chapter 4.4, the basic idea of constructing optimal path strategy is as follows: Let us start with an iterative scheme without the damping term. Then the strategy is constructed as follows: We fix a discretisation number N and perform enough number of iterations (the meaning of 'enough' will be specified in these following strategies). Next, we switch to a new discretisation number N and repeat this process for several different discretisation numbers. Finally, we perform 1 iteration after jumping to the final discretisation number N_f, which terminates the calculation. After giving some remarks, we will present our nearly optimal strategies for the linear and the nonlinear problem.

Remark 4.5.1.

- *In our previous calculations, we set the goal error as* $\varepsilon_N^k \leqslant \varepsilon_g = \frac{\varepsilon_{99}+\varepsilon_{100}}{2}$. *Here we recall that the error is defined as the energy difference* $\varepsilon_N^k = E(u_N^k) - E(u_f)$ *and* u_f, *in our case, is the converged solution for* $N = 200$, *i.e.,* $u_f \coloneqq u_{200}^{\infty}$. *However, in a more general case, we only know our approximate numerical solution* u_N^k. *Thanks to the* a posteriori *error estimation, the difference between the approximate solution and the unknown true solution* $\|u_N^k - u\|$ *is bounded by the residual under a certain norm. Therefore, in a more general situation, we set the criteria for the residual* $\|R(u_N^k)\| < \varepsilon_g$ *where* ε_g *here is given a priori (and not with respect to the knowledge of the error bound for* $N = 100$*).*

- *In the expression of the residual, the potential term* V *and the source term* f *in the operator* A *(in the linear case) or in the operator* F *(in the nonlinear case) are expressed as infinite sum of cosine functions. Even acting on an approximate solution* u_N^k, *which is a finite sum of cosine functions, the residual is still an infinite sum of cosine functions. Therefore, In our numerical simulation, we will numerically approximate the residual by truncating it for the sum of the first 1001 basis vectors* $(e_i)_{0\leqslant i\leqslant 1000}$ *as defined in* (4.9).

4.5.1 Strategy for linear problem

In this subsection, we will propose our strategies for the linear problem. In previous section, we analyze the mechanism producing two types of optimal paths. For **Type L** path, we know that the jump from small N to a larger N' is efficient because this iteration efficiency is controlled by the matrix norm $\|\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\|_{H^1\to H^1}$, which is smaller than $\|\mathbf{P}_{N'}\|_{H^1\to H^1}$. Therefore, let us start with proposing a calculation strategy based on the above mentioned mechanism in order to verify that we have well understood the optimal path behavior and that we are able to produce near-optimal path.

For a very well converged solution $\mathbf{u}_N^k \approx \mathbf{u}_N^{\infty}$, we can approximately predict the iteration error of the first iteration after jumping from N to N' via the following estimate:

$$\|\mathbf{u}_{N'}^1 - \mathbf{u}_{N'}^{\infty}\|_{H^1} \leqslant \|\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\|_{H^1\to H^1} \cdot \|\mathbf{u}_N^{\infty} - \mathbf{u}_{N'}^{\infty}\|_{H^1},$$

which means that we can control the error after the first iteration by calculating some terms independent of the iteration process. One application of this idea is as follows: We take $N' = 100$

as example and the goal is to obtain an approximate solution $\mathbf{u}_{100}^1$ (only one iteration for $N' = 100$ after the jump) such that $\|\mathbf{u}_{100}^1 - \mathbf{u}_{100}^\infty\|_{H^1} \leqslant \varepsilon_g$. Then we can determine the smallest N such that after the jump from $\mathbf{u}_N^k$ to $\mathbf{u}_{100}^0$, we only need one iteration to achieve the goal accuracy. This is realized by supposing that $\mathbf{u}_N^k$ is well converged to $\mathbf{u}_N^\infty$ and calculate the smallest N such that $\|\mathbf{P}_{100}(\mathbf{I} - \Pi_N)\|_{H^1 \to H^1} \cdot \|\mathbf{u}_N^\infty - \mathbf{u}_{100}^\infty\|_{H^1} \leqslant \varepsilon_g$.

We can also generalize this idea to design a nearly optimal path for achieving a target accuracy set to the iteration error $\|\mathbf{u}_{100}^k - \mathbf{u}_{100}^\infty\|_{H^1} \leqslant \varepsilon_g$. Here we remark that we design the path starting from $N_{\max} = 100$ and ending with $N_{\min} = 3$. For every jump, given the iteration error accuracy for N', we pick the smallest N such that after jumping from N to N' one iteration is enough to satisfy the required iteration error. In addition, we consider the difference between the numerical solution $\mathbf{u}_N^k$ and the converged solution $\mathbf{u}_N^\infty$. Therefore, let us denote by $\widetilde{\varepsilon_i}$ the iteration error tolerance to bound the iteration error, i.e., $\|\mathbf{u}_N^k - \mathbf{u}_N^\infty\|_{H^1} \leqslant \widetilde{\varepsilon_i}$ and *a priori*, this tolerance should be small. It follows that $\|\mathbf{u}_N^k - \mathbf{u}_{N'}^\infty\|_{H^1} \leqslant \|\mathbf{u}_N^\infty - \mathbf{u}_{N'}^\infty\|_{H^1} + \widetilde{\varepsilon_i}$. Now we start from $N = 100$ with the target accuracy $\|\mathbf{u}_{100}^1 - \mathbf{u}_{100}^\infty\|_{H^1} \leqslant \varepsilon_g$, we pick the smallest N such that $\|\mathbf{P}_{100}(\mathbf{I} - \Pi_N)\|_{H^1 \to H^1} \cdot (\|\mathbf{u}_N^\infty - \mathbf{u}_{100}^\infty\|_{H^1} + \widetilde{\varepsilon_i}) \leqslant \varepsilon_g$. Once again, we set the previous N before the jump as N' for the next loop and seek for a smaller N for the new jump. If there exist some N satisfying $\|\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\|_{H^1 \to H^1} \cdot (\|\mathbf{u}_N^\infty - \mathbf{u}_{N'}^\infty\|_{H^1} + \widetilde{\varepsilon_i}) \leqslant \widetilde{\varepsilon_i}$, then we pick the smallest value of N. If not, it means that the iteration error $\widetilde{\varepsilon_i}$ is not small enough. We add one GSR iteration for the fixed N' and reduce the target accuracy to $\|\mathbf{P}_{N'}\|_{H^1 \to H^1}\widetilde{\varepsilon_i}$. Then we check if we can pick a smaller N to realize a jump. If the answer is still no, we just repeat performing GSR iterations for the fixed N' until the target accuracy is small enough to realize a jump. Repeating the above process until, finally, we arrive at $N = N_{\min} = 3$.

In fact, even the above strategy produces near-optimal path that spends less computational resources compared with the optimal path obtained via the probabilistic calculation. This strategy is still expensive to implement: We must calculate the matrix norm $\|\mathbf{P}_{N'}(\mathbf{I} - \Pi_N)\|_{H^1 \to H^1}$ and discretisation error $\|\mathbf{u}_N^\infty - \mathbf{u}_{N'}^\infty\|_{H^1}$ for different N and N', which is more expensive than just solving directly the problem with fixed discretisation number $N = 100$. In addition, in practical calculation, we won't calculate $\mathbf{u}_N^\infty$ for different N and then set the iteration error value as target accuracy. Therefore, in our nearly optimal path strategy, we must avoid calculating the matrix norm for different N. Moreover, the target accuracy is set to the value of the residual $\|R(u_N^k)\| < \varepsilon_g$, which is more common in the general case.

In our strategy, the jump criteria is based on the use of the numerical solution residual in place of errors in H^1 norm, which is nearly free. Because our goal criteria is $\|R(u_N^k)\| < \varepsilon_g$ and we need to calculate the residual after each iteration to know if the target accuracy is achieved. After obtaining $\mathbf{u}_N^k$, we calculate its residual $R(u_N^k)$. By truncating the residual at N' ($N' > N$), we obtain

$$\begin{aligned}\Pi_{N'} R(u_N^k) &= \Pi_{N'}(-\Delta u_N^k + V u_N^k) - \Pi_{N'} f \\ &= \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k - \Pi_{N'} f,\end{aligned}$$

where A_1 and A_2 are defined through Corollary 4.1.1.1. Besides, for next iteration, if we jump from N to N' and perform one GSR iteration to obtain $u_{N'}^1$, then from (4.59), we have the following equation

$$\Pi_{N'} A_1 u_{N'}^1 + \Pi_{N'} A_2 u_N^k = \Pi_{N'} f.$$

Combing the above two equations yields that

$$\begin{aligned}\Pi_{N'}R(u_N^k) &= \Pi_{N'}A_1u_N^k + \Pi_{N'}A_2u_N^k - \Pi_{N'}f\\ &= \Pi_{N'}A_1u_N^k - \Pi_{N'}A_1u_{N'}^1\\ &= \Pi_{N'}A_1(u_N^k - u_{N'}^1).\end{aligned} \tag{4.148}$$

The above equation shows that by truncating the residual at N', this part could, in some sense, represent a substitute of $\|u_N^k - u_{N'}^1\|_{H^1}$ and predict the situation of the next iterative solution $u_{N'}^1$ after jumping from N to N'. Besides, for the rest part of the residual $\Pi_{N'}^{\perp}R(u_N^k)$, we have

$$\begin{aligned}\Pi_{N'}^{\perp}R(u_N^k) =&\Pi_{N'}^{\perp}(R(u_N^k) - R(u))\\ =&\Pi_{N'}^{\perp}A(u_N^k - u)\\ =&\Pi_{N'}^{\perp}A(u_N^k - u_{N'}^{\infty} + u_{N'}^{\infty} - u)\\ =&\Pi_{N'}^{\perp}A(u_N^k - u_{N'}^{\infty}) + \Pi_{N'}^{\perp}A(u_{N'}^{\infty} - u)\\ =&\Pi_{N'}^{\perp}A(u_N^k - u_{N'}^{\infty}) + R_{N'}^{\infty},\end{aligned} \tag{4.149}$$

where u is the true solution of this problem and we also used the fact that $u_{N'}^{\infty}$ is the discrete solution satisfying condition (4.60). Additionally, we have

$$\begin{aligned}\Pi_{N'}R(u_N^k) =&\Pi_{N'}(R(u_N^k) - R_{N'}^{\infty})\\ =&\Pi_{N'}A(u_N^k - u_{N'}^{\infty}).\end{aligned} \tag{4.150}$$

Therefore, if we assume that the term $\Pi_{N'}A(u_N^k - u_N^{\infty})$ is dominant in $A(u_N^k - u_N^{\infty})$, then the term $\Pi_{N'}^{\perp}R(u_N^k)$ can approximately represent $R_{N'}^{\infty}$. Here, we do one test: For $N = 3$, we calculate the residual R_N^{∞} and then for $N < N' \leqslant 100$, we compare $\|\Pi_{N'}^{\perp}R_N^{\infty}\|_{H^{-1}}$ and $\|R_{N'}^{\infty}\|_{H^{-1}}$. The result shows that for each $N < N' \leqslant 100$, we have $\|\Pi_{N'}^{\perp}R_N^{\infty}\|_{H^{-1}} > \|R_{N'}^{\infty}\|_{H^{-1}}$. This result shows that after getting a well-converged solution for $N = N_{\min}$, we can already predict N_f such that $R_{N_f}^{\infty} \leqslant \varepsilon_g$. The application of this character avoids picking unreasonably big discretisation number as the destination of jump.

After presenting all relevant formulas, now we propose the first nearly optimal path strategy in Algorithm 3. After giving a goal accuracy for the residual $\|R(u_{N_f}^k)\|_{H^{-1}} < \varepsilon_g$ (hopefully with $k = 1$), we begin from $N_{\min} = 3$ and make several jumps to obtain an approximate solution satisfying the target accuracy. Here, in the criteria $\|R_{\text{iter}}(u_N^k)\|_{H^{-1}} \geqslant \frac{\varepsilon_g}{10}$ and $\|\Pi_{N'}^{\perp}R(u_N^k)\|_{H^{-1}} \leqslant \frac{\|R_{\text{disc}}(u_N^k)\|_{H^{-1}}}{3}$, the value 10 and 3 are parameters to be fitted by experience. Besides, similar to the relation between $\Pi_{N'}R(u_N^k)$ and $R_{N'}^{\infty}$, we also perform the following numerical test: For a well converged solution u_N^k, by picking the smallest N' such that $\|\Pi_{N'}^{\perp}R(u_N^k)\|_{H^{-1}} \leqslant \frac{\|R_{\text{disc}}(u_N^k)\|_{H^{-1}}}{3}$, we perform one iteration after jumping from N to N' and calculate $\frac{\|R(u_N^k)\|_{H^{-1}}}{\|R_{N'}^1\|_{H^{-1}}}$. This ratio is very close to 3. In addition, the criteria for picking new N makes use of our previous observation about N_f. In addition, we also set restriction on the jump distance $N' - N$ to avoid picking clearly unreasonable values of N', e.g., in previous linear optimal path problem, performing iterations for $N = 300$ can satisfy the accuracy requirement but the computational cost will increase at least by a factor of 9 compare to the optimal path. Therefore, after each jump, we record the jump distance $N' - N$. When we obtain next N' through calculation, we compare the actual distance $N' - N$ with four times the previous one to avoid too big jump.

The inconvenient of the above strategy is that the criteria for picking new N contains two hand-fitted parameters (10 and 3). Therefore, we try in another strategy to give 'reasonable'

Algorithm 3 The nearly optimal path Strategy I

Input: target accuracy $\|Au_N^k\|_{H^{-1}} < \varepsilon_g$ and $N_{\min} = 3$.

$N = N_{\min}$
while $\|R(u_N^k)\|_{H^{-1}} > \varepsilon_g$ **do**
  while $\|R_{\text{iter}}(u_N^k)\|_{H^{-1}} \geqslant \frac{\varepsilon_g}{10}$ and $\|R(u_N^k)\|_{H^{-1}} > \varepsilon_g$ **do**
    One GSR iteration
  end while
  if $\|R(u_N^k)\|_{H^{-1}} \leqslant \varepsilon_g$ **then**
    break while iteration process
  end if
  Pick N_f as the smallest N' such that $\|\Pi_{N'}^{\perp} R(u_N^k)\|_{H^{-1}} \leqslant \varepsilon_g$
  Pick smallest N' such that $\|\Pi_{N'}^{\perp} R(u_N^k)\|_{H^{-1}} \leqslant \frac{\|R_{\text{disc}}(u_N^k)\|_{H^{-1}}}{3}$
  Pick $N^* = \min\{N_f, N'\}$
  if $N^* - N$ is four times bigger than last jump **then**
    $N \leftarrow \frac{N^*+N}{2}$
  else
    $N \leftarrow N^*$
  end if
end while

Output: Nearly optimal path S and the approximate solution u_N^k.

criteria for picking next N'. We firstly divide the strategy into two parts: the first part containing all the jumps except the last one jumping to the final discretisation number N_f and the other part only containing the last jump. We make this division because these two parts serve different purposes. According to the expression of cost function (4.140), for the aim of decreasing computational costs, we should decrease the number of iterations for big discretisation number, e.g. only 1 iteration for $N = N_f$. Therefore, all these rest iterations aim at decreasing the error (or the residual) at a certain level such that 1 iteration for $N = N_f$ is enough to finalize the calculation.

For the first part of the path, we want to decrease the error while balancing corresponding computation costs. Therefore, after obtaining a well-converged solution u_N^k, we define a efficiency parameter $\kappa_{N'}$ to evaluate the efficiency if we pick discretisation number N' as the destination of the jump. Then we pick N' corresponding to the highest value of the efficiency parameter κ. Ideally, the efficiency parameter should be defined as $\kappa_{N'} := \frac{\Delta R}{\Delta C}$, where ΔR is the variation of residual after doing $K_{N'}$ iterations for N' such that the iteration residual $R_{\text{iter}}(u_{N'}^{K_{N'}})$ satisfies the stopping criteria and $\Delta C = K_{N'}(N'+1)^2$ is the increase of computation cost. Nevertheless, as we said before, it isn't practical to calculate $\kappa_{N'}$ precisely, which is as expensive as solving the problem directly for fixed $N = 100$. Therefore, we need to find a computationally cheap approximation of this efficiency parameter.

The practical efficiency parameter is defined as

$$\kappa_{N'} = \frac{\|R(u_N^k)\|_{H^{-1}} - \|\Pi_{N'}^{\perp} R(u_N^k)\|_{H^{-1}}}{(N'+1)^2}, \tag{4.151}$$

where u_N^k is the actual numerical solution before jumping. This new definition bases on some numerical observations.

The first observation is that the spectral radius of $\mathbf{P}_N$ is nearly a constant with respect to N.

In fact, theoretically, the spectral radius of $\mathbf{P}_N$ converges to the operator norm of $A_1^{-1}A_2$ and our numerical test shows that the values of $\mathbf{P}_N$ for different N don't vary much, even for small N. Moreover, recall that the decrease rate of iteration error is closer and closer to the spectral radius of $\mathbf{P}_N$ as the iteration process goes on. So is the iteration residual according to the iteration residual definition (4.66). Therefore, we can store the iteration residual values for $N_{\text{ini}} = 3$. If we could predict the iteration residual of $u^1_{N'}$ after jumping from N to N' and performing 1 iteration, then we could predict the number of iterations $K_{N'}$ needed for the new discretisation number N'. Therefore, we did a test: For $N = 5$, we take u^∞_N as input. Then for different $N < N' \leqslant 100$, we jump from N to N' and calculate output $u^1_{N'}$ and the corresponding iteration residual $R_{\text{iter}}(u^1_{N'})$. We show the result in Figure 4.12. From Figure 4.12, we observe that the iteration residual increases for N' close to initial input N. Then it is nearly a constant for other choice of output discretisation number N'. Combining the above arguments, if we suppose that the discretisation number corresponding to the best efficiency locates in the constant iteration residual value region, then the number of iterations $K_{N'}$ is approximately regarded as a constant for different N'. This explains why we remove the term $K_{N'}$ in the denominator of $\kappa_{N'}$ in (4.151). For the term ΔR in (4.151), $\|\Pi^\perp_{N'} R(u^k_N)\|_{H^{-1}}$ is used to approximately represent $\|R(u^{K_{N'}}_{N'})\|_{H^{-1}}$ as discussed previously.

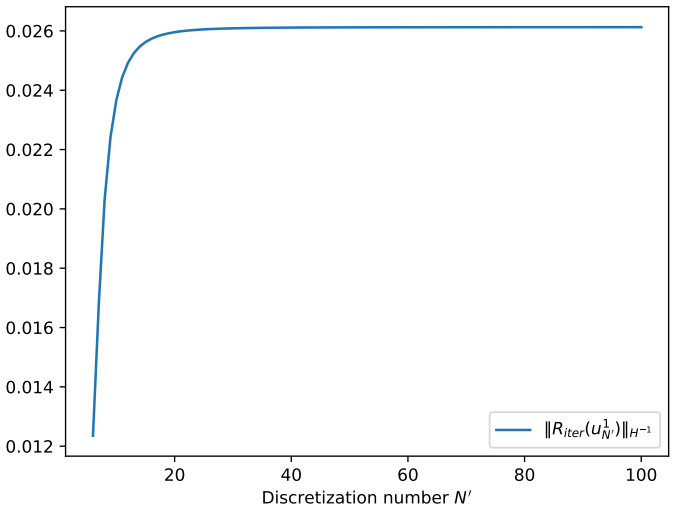


Figure 4.12: Iteration residual of $u^1_{N'}$ for different discretisation number N'.

For the second part of the optimal path, namely, the last jump to N_f, we make use of the experience result. We estimate the iteration residual and the discretisation residual of $u^1_{N_f}$ if we jump from N to N_f and perform 1 GSR iteration. Then we jump for N large enough such that the sum of the estimated iteration and discretisation residuals are smaller than the target accuracy. After obtaining u^k_N, if we jump from N to $N_f = 100$ and perform the last iteration, the estimation of the discretisation residual of $u^1_{N_f}$ is $\|\Pi^\perp_{N_f} R(u^k_N)\|_{H^{-1}}$. However, we could not connect directly the term $A_2(u^1_{N_f} - u^k_N)$ with $R(u^k_N)$. From the definition of operator A_2 in Proposition 4.1.2, it is easy to deduce that the regularity of A_2 is the same as that of the potential function V, i.e., $A_2 \in L^\infty(\mathbb{T}) \cap H^{\frac{3}{2}-\epsilon}$ $(\epsilon > 0)$ for our chosen V in Remark 4.1.2. Thus

it follows that

$$\begin{aligned}
\|A_2(u^1_{N_f}-u^k_N)\|_{H^{-1}} &= \sup_{v\in X, v\neq 0} \frac{\left\langle A_2(u^1_{N_f}-u^k_N), v\right\rangle_{X',X}}{\|v\|_{H^1}} \\
&= \sup_{v\in X, v\neq 0} \frac{\int_0^{2\pi} A_2(u^1_{N_f}-u^k_N)v}{\|v\|_{H^1}} \\
&= \sup_{v\in X, v\neq 0} \frac{\int_0^{2\pi} (u^1_{N_f}-u^k_N)(A_2 v)}{\|v\|_{H^1}} \\
&\leqslant \sup_{v\in X, v\neq 0} \frac{\|A_2 v\|_{H^1}\ \|u^1_{N_f}-u^k_N\|_{H^{-1}}}{\|v\|_{H^1}} \\
&\leqslant \sup_{v\in X, v\neq 0} \frac{\|A_2\|_{H^1}\ \|v\|_{H^1}\ \|u^1_{N_f}-u^k_N\|_{H^{-1}}}{\|v\|_{H^1}} \\
&\leqslant \|A_2\|_{H^1}\ \|u^1_{N_f}-u^k_N\|_{H^{-1}}.
\end{aligned} \tag{4.152}$$

From the super convergence result in Theorem 4.1.1, we can deduce that $u^1_{N_f}-u^k_N \approx (\mathbf{I}-\Pi_N)u^1_{N_f}$. Thus similarly to Estimate (4.19), we have

$$\|u^1_{N_f}-u^k_N\|_{H^{-1}} \leqslant \frac{C_r}{N^2}\|u^1_{N_f}-u^k_N\|_{H^1}, \tag{4.153}$$

where $C_r > 1$ is close to 1 for N and N_f large enough. Lastly, assuming that the Laplace operator $-\Delta$ is dominant in A_1, we have

$$\|A_1(u^1_{N_f}-u^k_N)\|_{H^{-1}} \approx \|u^1_{N_f}-u^k_N\|_{H^1}. \tag{4.154}$$

Combining all these above arguments, we approximately have

$$\begin{aligned}
\|R_{\text{iter}}(u^1_{N_f})\|_{H^{-1}} &\leqslant \|A_2\|_{H^1}\ \|u^1_{N_f}-u^k_N\|_{H^{-1}} \\
&\leqslant \frac{C_r\|A_2\|_{H^1}}{N^2}\|u^1_{N_f}-u^k_N\|_{H^1} \\
&\approx \frac{C_r\|A_2\|_{H^1}}{N^2}\|A_1(u^1_{N_f}-u^k_N)\|_{H^1}.
\end{aligned} \tag{4.155}$$

Based on the above estimate, we perform the following numerical test: We fix $N_f = 100$ and take u^∞_N as input for $3 \leqslant N \leqslant N_f - 1$. Then we jump from N to N_f and perform 1 GSR iteration to get $u^1_{N_f}$ and $\|R_{\text{iter}}(u^1_{N_f})\|_{H^{-1}}$. We plot the variation of $\|R_{\text{iter}}(u^1_{N_f})\|_{H^{-1}}$ and $\frac{\|\Pi_{N_f}R(u^\infty_N)\|_{H^{-1}}}{N^2}$ in Figure 4.13. From this figure, it might be reasonable to use $\frac{\|\Pi_{N_f}R(u^k_N)\|_{H^{-1}}}{N^2}$ as the estimation of the iteration residual in our strategy. Combining all these above statements, we propose our second nearly optimal strategy in Algorithm 4.

After giving our strategies, we do a comparison with the optimal path. Recalling that In the optimal path problem, the target error is the energy difference $\varepsilon^k_N \leqslant \varepsilon_g = \frac{\varepsilon^\infty_{99}+\varepsilon^\infty_{100}}{2}$. We need to transfer it into the residual requirement: Firstly, we fix $N = 100$ and get u^k_{100} such that $\varepsilon^k_{100} \leqslant \varepsilon_g$ after performing the smallest number of iterations. Then we set $\|R(u^k_{100})\|_{H^{-1}}$ as the target accuracy. Finally, we get our nearly optimal strategies. Together with the optimal path and the path for fixed $N = 100$, we plot them in Figure 4.14 and list corresponding computational

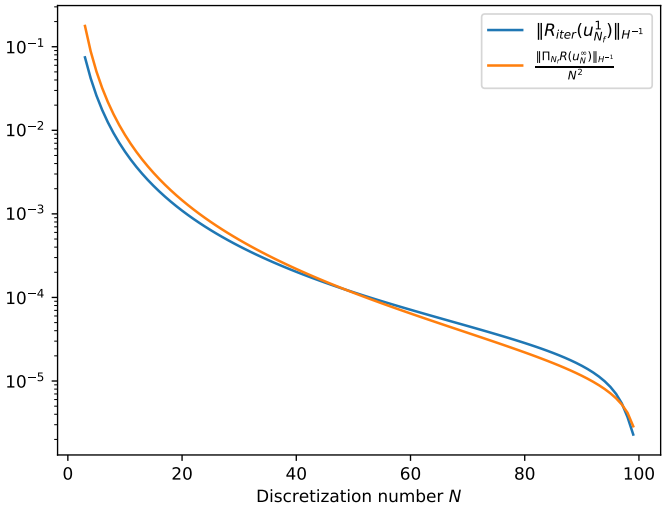


Figure 4.13: Iteration residual of u_N^1 for different discretisation number N.

costs in Table 4.3.

Path	Fixed $N = 100$	Optimal path	Strategy I	Strategy II
Cost	132,613	10,462	13,126	12,136

Table 4.3: Computational costs of different paths for the linear problem.

Compared to the optimal path $\text{Cost}(S_{\text{op}}) = 10,462$, Strategy I offers an nearly optimal path with $\text{Cost}(S_{\text{I}}) = 13,126$. The main contribution of the cost comes from the iteration for $N = 105$. From this result, we know that the jump in Strategy I is still a little large and that it could not predict N_f precisely. Besides, we have 2 iterations for $N = 28$, which makes our path a bit more expensive than the optimal one. Nevertheless, the nearly optimal path is finalized with only 1 iteration for the largest N and the cost isn't far from the optimal one. Strategy II offers an nearly optimal path with a lower cost $\text{Cost}(S_{\text{II}}) = 12,136$. The second strategy gives a more accurate prediction, i.e., $N_f = 101$. In addition, we have more iterations for small N in Strategy II. However, even we only count these computational costs originating from small N and compare these costs for strategy I and II, Strategy II still offers a more optimal path. From this comparison, we know that when we have less iterations for bigger N, we have a more optimal path. In conclusion, this result shows that both of our Strategy I and II offer nearly optimal paths well balancing between the cost and accuracy.

4.5.2 Strategy for the nonlinear problem

In this section, we propose the nonlinear calculation strategy, which is similar to that for the linear problem. For our nonlinear problem, we have two iterative schemes at hand with different characters. For the scheme without the damping term, the nonlinear term $\mathbf{S}_N$ is constructed from $\frac{\mathbf{u}_N^k + \mathbf{u}_N^{k-1}}{2}$ but not $\mathbf{u}_N^k$. For fixed N, the convergence of the iteration process is assured.

Algorithm 4 The nearly optimal path Strategy II

Input: target accuracy $\|Au_N^k\|_{H^{-1}} < \varepsilon_g$ and $N_{\min} = 3$.

$N = N_{\min}$
while $\|R(u_N^k)\|_{H^{-1}} > \varepsilon_g$ **do**
  while $\|R_{\text{iter}}(u_N^k)\|_{H^{-1}} \geqslant \frac{\varepsilon_g}{10}$ and $\|R(u_N^k)\|_{H^{-1}} > \varepsilon_g$ **do**
    One GSR iteration
  end while
  if $\|R(u_N^k)\|_{H^{-1}} \leqslant \varepsilon_g$ **then**
    break while iteration process
  end if
  Pick N_f as the smallest N' such that $\|\Pi_{N'}^{\perp} R(u_N^k)\|_{H^{-1}} \leqslant \varepsilon_g$
  if $\|\Pi_{N_f}^{\perp} R(u_N^k)\|_{H^{-1}} + \frac{\|\Pi_{N_f} R(u_N^k)\|_{H^{-1}}}{N^2} \leqslant \varepsilon_g$ **then**
    $N \leftarrow N_f$
  else
    Calculate $\kappa_{N'} = \frac{\|R(u_N^k)\|_{H^{-1}} - \|\Pi_{N'}^{\perp} R(u_N^k)\|_{H^{-1}}}{(N'+1)^2}$ for $N < N' < N_f$ and pick N' corresponding to the biggest $\kappa_{N'}$.
    $N \leftarrow N'$
  end if
end while

Output: Nearly optimal path S and the approximate solution u_N^k.

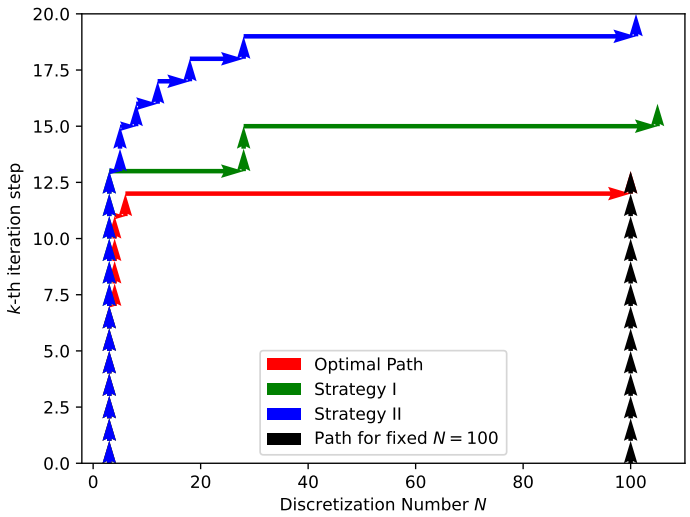


Figure 4.14: Comparison between optimal path and nearly optimal strategies

With this scheme, we have the **Type L** optimized path result shown in Figure 4.11. For the scheme with damping term, the convergence of the iteration process for fixed N is also assured. However, in this situation, there is no efficient jump as explained earlier.

Here we propose two associated strategies for the nonlinear problem. The Strategy I is associated with the scheme of **Type L** (without the damping term but with the nonlinear term

$\mathbf{S}_N$ constructed from $\frac{\mathbf{u}_N^k+\mathbf{u}_N^{k-1}}{2}$). The Strategy II is associated with the scheme of **Type NL** (with damping term for iterations with fixed N) but without the damping term for the first iteration after a jump. For each of those two schemes, the convergence of the solution is achieved while the jump efficiency is maintained.

Since we propose iterative scheme (4.147) when performing complementary tests in Chapter 4.4, the *a posteriori* error analysis of this iterative scheme is undertaken. Here before applying our strategies to these two iterative schemes, we firstly provide the following *a posteriori* error analysis of iterative scheme (4.147).

According to the expression of the iterative scheme given by Equation (4.147), we define the discretisation residual as follows

$$R_{\text{disc}}(u_N^k) \coloneqq A_1 u_N^k + A_2 u_N^{k-1} + \left(\frac{u_N^{k-1}+u_N^{k-2}}{2}\right)^2 u_N^{k-1} - f, \tag{4.156}$$

such that compared to the linear case, we maintain the property that

$$\Pi_N R_{\text{disc}}(u_N^k) = 0, \tag{4.157}$$

where A_1 and A_2 are defined in Corollary 4.1.1.1 and we denote by $A = A_1 + A_2$. Then we define the iteration residual as the rest part in the total residual

$$\begin{aligned}
R_{\text{iter}}(u_N^k) \coloneqq & R(u_N^k) - R_{\text{disc}}(u_N^k) \\
= & A_1 u_N^k + A_2 u_N^k + (u_N^k)^3 - f - A_1 u_N^k - A_2 u_N^{k-1} - \left(\frac{u_N^{k-1}+u_N^{k-2}}{2}\right)^2 u_N^{k-1} + f \\
= & A_2(u_N^k - u_N^{k-1}) + (u_N^k)^3 - \left(\frac{u_N^{k-1}+u_N^{k-2}}{2}\right)^2 u_N^{k-1} \\
= & A_2(u_N^k - u_N^{k-1}) + (u_N^k)^3 - (u_N^{k-1})^3 + (u_N^{k-1})^3 - \left(\frac{u_N^{k-1}+u_N^{k-2}}{2}\right)^2 u_N^{k-1} \\
= & A_2(u_N^k - u_N^{k-1}) + (((u_N^k)^2 + (u_N^{k-1})^2 + u_N^k u_N^{k-1})(u_N^k - u_N^{k-1}) \\
& + \frac{1}{4} u_N^{k-1}(3u_N^{k-1} + u_N^{k-2})(u_N^{k-1} - u_N^{k-2}) \\
& .
\end{aligned} \tag{4.158}$$

Now we state the lemma showing relation between error and residual.

Lemma 4.5.1. *For $N \in \mathbb{N}^*$, let the discretisation residual be defined through* (4.156)*, let the iteration residual be defined through* (4.158) *and let the iteration error and the discretisation error be defined through* (4.61)*. Then*

- *The iteration residual is bounded above by the iteration error:*

$$\begin{aligned}
\|R_{\text{iter}}(u_N^k)\|_{H^{-1}} \leqslant & \|A_2\|_{H^1 \to H^{-1}} \left(\text{Er}_{\text{iter}}(u_N^k) + \text{Er}_{\text{iter}}(u_N^{k-1})\right) \\
& + C_{GN}^2(\|u_N^k\|_{L^2}^2 + \|u_N^k\|_{L^2}\|u_N^{k-1}\|_{L^2} + \|u_N^{k-1}\|_{L^2}^2)\left(\text{Er}_{\text{iter}}(u_N^k) + \text{Er}_{\text{iter}}(u_N^{k-1})\right) \\
& + \frac{1}{4} C_{GN}^2 \|u_N^{k-1}\|_{L^2}(3\|u_N^{k-1}\|_{L^2} + \|u_N^{k-2}\|_{L^2})\left(\text{Er}_{\text{iter}}(u_N^{k-1}) + \text{Er}_{\text{iter}}(u_N^{k-2})\right),
\end{aligned} \tag{4.159}$$

where C_{GN} is the Gagliardo Nirenberg type inequality constant.

- *The discretisation residual is bounded above by the discretisation error and the iteration error:*

$$\begin{aligned}\|R_{\text{disc}}(u_N^k)\|_{H^{-1}} \leqslant & \|A_1\|_{H^1\to H^{-1}}\text{Er}_{\text{iter}}(u_N^k) + \|A_2\|_{H^1\to H^{-1}}\text{Er}_{\text{iter}}(u_N^{k-1}) + \beta_a \text{Er}_{\text{disc}} \\ & + C_{GN}^2\left(\text{Er}_{\text{iter}}(u_N^{k-1}) + \text{Er}_{\text{disc}}\right)\left(3\|u_N^{k-1}\|_{L^2}^2\right. \\ & +3\|u_N^{k-1}\|_{L^2}\left(\text{Er}_{\text{iter}}(u_N^{k-1}) + \text{Er}_{\text{disc}}\right) \\ & \left. + \left(\text{Er}_{\text{iter}}(u_N^{k-1}) + \text{Er}_{\text{disc}}\right)^2\right) \\ & + \frac{1}{4}C_{GN}^2\|u_N^{k-1}\|_{L^2}(3\|u_N^{k-1}\|_{L^2} + \|u_N^{k-2}\|_{L^2})\left(\text{Er}_{\text{iter}}(u_N^{k-1}) + \text{Er}_{\text{iter}}(u_N^{k-2})\right),\end{aligned} \tag{4.160}$$

where β_a is the continuity constant of $a\colon X\times X\to\mathbb{R}$ defined through (4.6). Besides, when the iteration error is small enough, the discretisation residual is mainly bounded by the discretisation error.

Proof. The proof is similar to that of the Lemma 4.2.4. For the iteration residual, we firstly rewrite

$$\begin{aligned}R_{\text{iter}}(u_N^k) =& A_2(u_N^k - u_N^{k-1}) + (((u_N^k)^2 + (u_N^{k-1})^2 + u_N^k u_N^{k-1})(u_N^k - u_N^{k-1}) \\ & + \frac{1}{4}u_N^{k-1}(3u_N^{k-1} + u_N^{k-2})(u_N^{k-1} - u_N^{k-2}) \\ =& A_2(u_N^k - u_N) + A_2(u_N - u_N^{k-1}) + (((u_N^k)^2 + (u_N^{k-1})^2 + u_N^k u_N^{k-1})(u_N^k - u_N) \\ & + (((u_N^k)^2 + (u_N^{k-1})^2 + u_N^k u_N^{k-1})(u_N - u_N^{k-1}) + \frac{1}{4}u_N^{k-1}(3u_N^{k-1} + u_N^{k-2})(u_N^{k-1} - u_N) \\ & + \frac{1}{4}u_N^{k-1}(3u_N^{k-1} + u_N^{k-2})(u_N - u_N^{k-2}).\end{aligned} \tag{4.161}$$

Following similar argument used to establish Estimate (4.130) and combining it with the Equation (4.161) yields the Estimate (4.159).

For the discretisation residual. We have

$$\begin{aligned}R_{\text{disc}}(u_N^k) =& A_1u_N^k + A_2u_N^{k-1} + \left(\frac{u_N^{k-1} + u_N^{k-2}}{2}\right)^2 u_N^{k-1} - Au - u^3 \\ =& A_1(u_N^k - u_N) + A_2(u_N^{k-1} - u_N) + A(u_N - u) + (((u)^2 + (u_N^{k-1})^2 + uu_N^{k-1})(u_N^{k-1} - u) \\ & + \frac{1}{4}u_N^{k-1}(3u_N^{k-1} + u_N^{k-2})(u_N^{k-2} - u_N^{k-1}) \\ =& A_1(u_N^k - u_N) + A_2(u_N^{k-1} - u_N) + A(u_N - u) \\ & + (3(u_N^{k-1})^2 - 3u_N^{k-1}(u_N^{k-1} - u) + (u_N^{k-1} - u)^2)(u_N^{k-1} - u) \\ & + \frac{1}{4}u_N^{k-1}(3u_N^{k-1} + u_N^{k-2})(u_N^{k-2} - u_N^{k-1}).\end{aligned} \tag{4.162}$$

Using similar argument as what we do in (4.132), (4.133) and (4.134), we obtain Estimate (4.160). When k goes to infinity, the iteration errors $\text{Er}_{\text{iter}}(u_N^{k-2})$, $\text{Er}_{\text{iter}}(u_N^{k-1})$ and $\text{Er}_{\text{iter}}(u_N^k)$ go to zero. Therefore the contribution of the discretisation residual is dominant in the above estimate when k is large enough. □

By truncating the residual, we have the following expression

$$\Pi_{N'} R(u_N^k) = \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^3 - \Pi_{N'} f,$$

where A_1 and A_2 are defined in (4.59). For the first scheme with the nonlinear term $\frac{\mathbf{u}_N^k + \mathbf{u}_N^{k-1}}{2}$, we have

$$\Pi_{N'} A_1 u_{N'}^1 + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (\frac{u_N^k + u_N^{k-1}}{2})^2 u_N^k = \Pi_{N'} f. \tag{4.163}$$

By assuming that the iteration error for u_N^k is small such that we have $u_N^k \approx u_N^{k-1}$, we approximately have

$$\begin{aligned} \Pi_{N'} R(u_N^k) &= \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^3 - \Pi_{N'} f \\ &\approx \Pi_{N'} A_1 u_N^k - \Pi_{N'} A_1 u_{N'}^1 \\ &\approx \Pi_{N'} A_1 (u_N^k - u_{N'}^1). \end{aligned} \tag{4.164}$$

For the second scheme, there is no damping term for the first iteration after the jump. So we have

$$\Pi_{N'} A_1 u_{N'}^1 + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^2 u_N^k = \Pi_{N'} f.$$

Combine with the expression of the projected residual (4.163), we still have

$$\begin{aligned} \Pi_{N'} R(u_N^k) &= \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^3 - \Pi_{N'} f \\ &= \Pi_{N'} A_1 (u_N^k - u_{N'}^1). \end{aligned} \tag{4.165}$$

On the other hand, for iterations with damping term (fixed discretisation number N), we have

$$\Pi_N A_1 u_N^{k+1} - (1-\xi) \Pi_N A_1 u_N^k + \xi \Pi_N A_2 u_N^k + \xi \Pi_N (u_N^k)^3 = \xi \Pi_N f,$$

or

$$\frac{1}{\xi} \Pi_N A_1 u_N^{k+1} - \frac{1-\xi}{\xi} \Pi_N A_1 u_N^k + \Pi_N A_2 u_N^k + \Pi_N (u_N^k)^2 u_N^k = \Pi_N f.$$

If we apply the above scheme to the first iteration after the jump (from N to N'), we have

$$\begin{aligned} \Pi_{N'} R(u_N^k) =& \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^3 - \Pi_{N'} f \\ =& \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^3 \\ & - (\frac{1}{\xi} \Pi_{N'} A_1 u_{N'}^1 - \frac{1-\xi}{\xi} \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^3) \\ =& \frac{1}{\xi} \Pi_{N'} A_1 (u_N^k - u_{N'}^1). \end{aligned} \tag{4.166}$$

From the above derivation, we deduce that for the first iteration after a jump, the difference $u_{N'}^1 - u_N^k$ depends on the choice of scheme (with or without damping term) and there is a ratio $\frac{1}{\xi}$ between these two cases. This also explains the fact that the first iteration after a jump loss its effectiveness for the scheme with damping term.

Moreover, in the nonlinear problem, we also test that for $N = 3$ and $N < N' \leqslant 100$, we always have $\|\Pi_{N'}^{\perp} R_N^{\infty}\|_{H^{-1}} > \|R_{N'}^{\infty}\|_{H^{-1}}$. Therefore, in the nonlinear problem, we can also predict N_f such that $\|R_{N_f}^{\infty}\|_{H^{-1}} \leqslant \varepsilon_g$. Additionally, for the iteration residual after the jump, the numerical observation shown in the linear case also holds except that the parameter connecting

the discretisation residual and the iteration residual is not N^2 but $N^{1.6}$ (obtained by trial and error and corresponding simulation result is shown in Figure 4.15), this difference is caused by the addition of nonlinear term. Nevertheless, even there is the nonlinear term, the behavior of the nonlinear problem is similar to that for the linear case and we therefore propose these same strategies as those in the linear case, in Algorithm 3 and in Algorithm 4, except that the iteration cost in efficiency parameter (4.151) becomes $(N+1)^2(N+2)$. Finally, we get these two nearly optimal paths in Figure 4.16 for the first scheme and Figure 4.17 for the second scheme and we list all these computational costs in Table 4.4. Similar to the linear problem, for two iterative schemes conserving the jump efficiency, our two proposed strategies can produce satisfactory nearly optimal paths. In addition, from Figure 4.15, we know that we overestimate the iteration residual for the first iteration after jumping from N to $N_f = 100$. Therefore, the prediction of N_f is also guaranteed. After jumping to $N_f = 100$, one iteration is enough to terminate the path.

Path	Fixed $N = 100$	Strategy I	Strategy II
Cost of Scheme I	16,648,032	1,279,222	1,150,010
Cost of Scheme II	15,607,530	1,354,672	1,150,070

Table 4.4: Computational costs of different paths for the nonlinear problem.

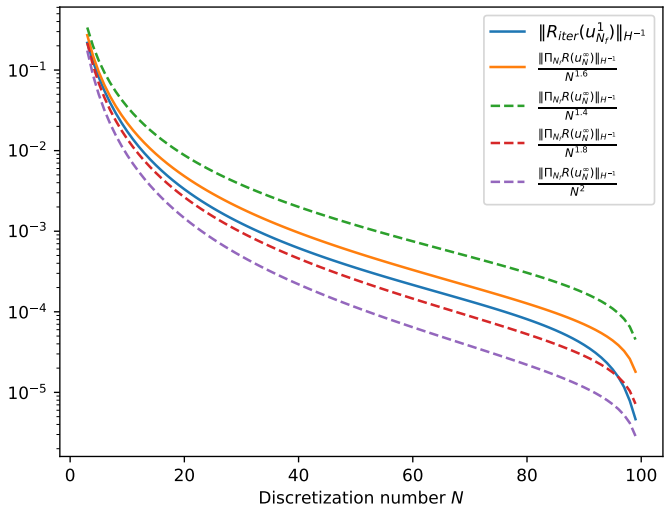


Figure 4.15: Iteration residual of u_N^1 for different discretisation number N.

4.6 Generalization of the problem

In previous sections, we study the optimal path problem and analysis the mechanism of the jump efficiency, which is the key feature determining the optimal path. Then with the help of the a posteriori error estimation and the residual decomposition, we propose two nearly optimal strategies, both of which produce satisfactory nearly optimal paths. In this section, we will

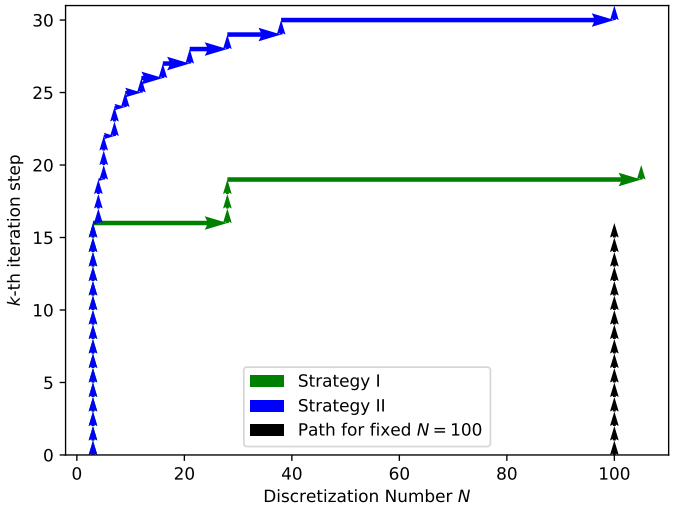


Figure 4.16: Nearly optimal strategies of the first scheme for the nonlinear case

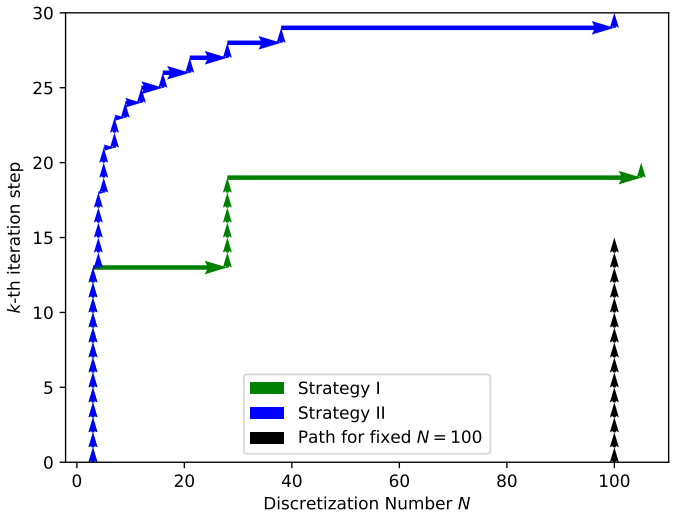


Figure 4.17: Nearly optimal strategies of the second scheme for the nonlinear case

change the regularity or continuity of the potential function V or the source term f in previous linear problem (4.2) and nonlinear problem (4.80). The aim of this section is to see if those two nearly optimal strategies still behave well under such modifications.

Recall that in the linear problem (4.2) and nonlinear problem (4.80), we pick

$$\forall x \in [0, 2\pi], \quad V(x) = 1 + \sum_{i \in \mathbb{N}^*} \frac{\cos(ix)}{i^2} \quad \text{and} \quad f(x) = \sum_{i \in \mathbb{N}^*} \frac{2\cos(ix)}{|i|^{0.05}}.$$

The first series of cases begin by setting

$$\forall x \in [0, 2\pi], \quad V(x) = 1 + \sum_{i \in \mathbb{N}^*} \frac{\cos(ix)}{i^\upsilon} \quad \text{and} \quad f(x) = \sum_{i \in \mathbb{N}^*} \frac{2\cos(ix)}{|i|^\tau}, \tag{4.167}$$

where $\upsilon = 1.5, 2.0$ or 2.5 and τ pick one of the following values $\{0.01,\ 0.05,\ 0.25,\ 0.5,\ 1.0\}$. And we still use the same iterative scheme for solving the linear or the nonlinear problem with this new series of parameters.

Similarly as in the optimal path problem, at the beginning the target accuracy is defined as the energy difference $\varepsilon_N^k \leqslant \varepsilon_g = \frac{\varepsilon_{99}^\infty + \varepsilon_{100}^\infty}{2}$ where ε_N^∞ $(3 \leqslant N \leqslant 100)$ is defined in (4.137). Then we transfer it into the residual requirement: We fix $N = 100$ and get u_{100}^k such that $\varepsilon_{100}^k \leqslant \varepsilon_g$ and that k is the smallest number of iterations. Then we set $\|R_{100}^k\|_{H^{-1}}$ as the target accuracy. Finally, we get our nearly optimal paths.

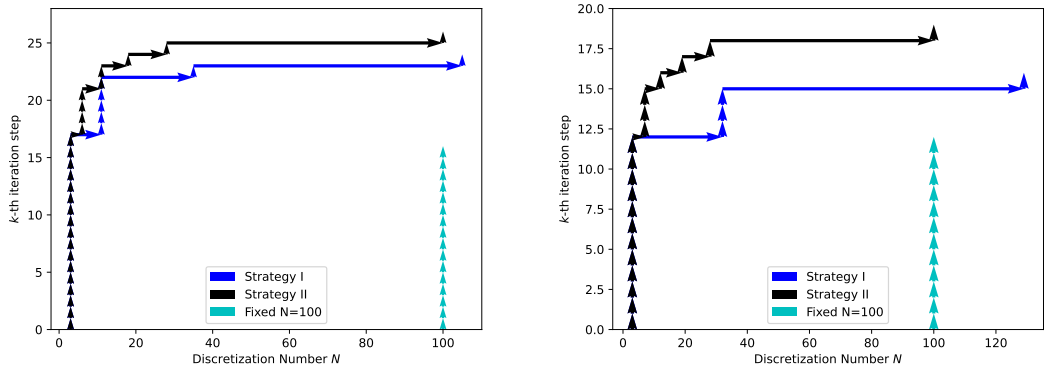
We calculate all above cases with different parameters υ and τ for the linear problem (4.2) with GSR iterative scheme and the nonlinear problem (4.80) with two iterative schemes mentioned in Chapter 4.5.2. The behavior of nearly optimal paths for different cases is quite similar: for each case, the Strategy II can predict accurately the final jump such that N_f is close to the theoretically smallest one $N = 100$. However, in some cases, the Strategy I can't do that, i.e., sometimes the final jump can reach $N \approx 130$. The behavior of these optimal paths is the same for both the linear problem and the nonlinear problem with two different iterative schemes. Therefore, here we just show near-optimal paths with two different choices of parameters for the linear problem in Figure 4.18. Figure 4.18A is the calculation result for $\upsilon = 2.5$ and $\tau = 0.5$ and Figure 4.18B is the calculation result for $\upsilon = 2.0$ and $\tau = 0.01$ with corresponding computational costs listed in Table 4.5.

Path6	Fixed $N = 100$	Strategy I	Strategy II
Cost with $\upsilon = 2.5$ and $\tau = 0.5$	163,216	13,524	12,159
Cost with $\upsilon = 2.0$ and $\tau = 0.01$	122,412	20,359	11,995

Table 4.5: Computational costs of different paths for different parameter values.

In above cases, the regularity of the solution u varies with different values of υ and τ with the variation of its Fourier coefficients $(\hat{u}_i)i \in \mathbb{N}$ being regular. In the following test, we introduce piecewise regularity variations of Fourier coefficients and verify whether our strategies can detect the connection points. The second series of cases begin by giving the explicit expression of the discrete solution $u_{1000} = \sum_{k=0}^{1000} \hat{u}_k e_k$:

$$\hat{u}_k = \begin{cases} \frac{2}{|20.5|}, & \text{if } k = 0 \\ \frac{2}{|k-20.5|} + \frac{2}{|k+20.5|}, & \text{if } 1 \leqslant k \leqslant 40 \\ \frac{2}{k^3}, & \text{if } k \geqslant 41 \end{cases}, \tag{4.168}$$



(A) Nearly optimal strategies with $\upsilon = 2.5$ and $\tau = 0.5$ for the linear problem.

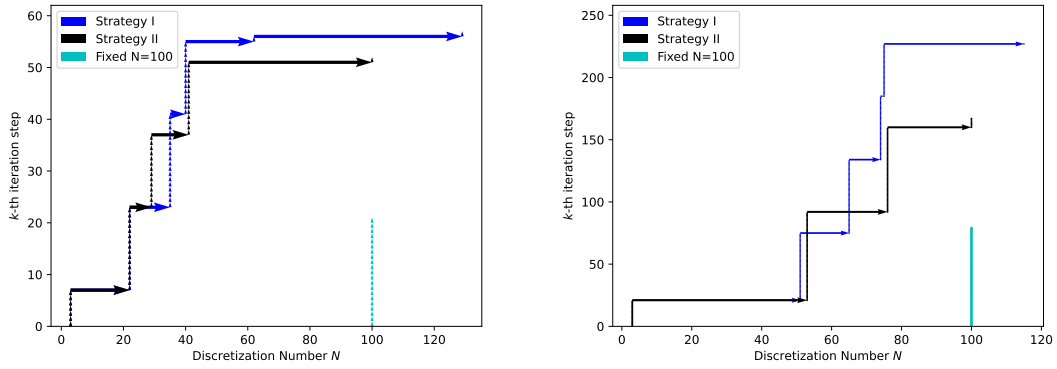
(B) Nearly optimal strategies with $\upsilon = 2.0$ and $\tau = 0.01$ for the linear problem.

Figure 4.18: Examples of nearly optimal strategies results.

or

$$\hat{u}_k = \begin{cases} \frac{2}{|25.5|} + \frac{2}{|25.5|}, & \text{if } k = 0 \\ \frac{2}{|k-50.5|} + \frac{2}{|k-25.5|} + \frac{2}{|k+25.5|} + + \frac{2}{|k+50.5|}, & \text{if } 1 \leqslant k \leqslant 75 \\ \frac{2}{k^3}, & \text{if } k \geqslant 76 \end{cases} . \tag{4.169}$$

By inserting the solution with expression (4.168) or (4.169) into the linear problem (4.14) with the same potential function V, we obtain the source term f'. Then we solve the new linear problem with new source term f'. Similarly, following the same manner, we can resolve new nonlinear problem by computing f' with (4.168) or (4.169). In addition, by varying the parameter υ in the potential function V, we get new series of source term f'.



(A) Nearly optimal strategies with $\upsilon = 1.5$ and solution (4.168) for the nonlinear problem.

(B) Nearly optimal strategies with $\upsilon = 2.5$ and solution (4.169) for the nonlinear problem.

Figure 4.19: Examples of nearly optimal strategies results.

We calculate all above cases with different parameters υ and f' for the linear problem (4.2)

Path6	Fixed $N = 100$	Strategy I	Strategy II
Cost with $\upsilon = 1.5$ and solution (4.168)	21,850,542	4,523,176	2,696,726
Cost with $\upsilon = 2.5$ and solution (4.169)	83,240,160	67,016,232	51,160,092

Table 4.6: Computational costs of different paths for different parameter values.

with GSR iterative scheme and nonlinear problem (4.80) with two iterative schemes mentioned in Chapter 4.5.2. The behavior of nearly optimal paths for different cases is similar to that for the first series of calculations: the Strategy II can predict accurately the final jump such that N_f is close to the theoretically smallest one $N = 100$. However, in some cases, the Strategy I can't do that. In addition, the behavior of these optimal paths is the same for both the linear problem and the nonlinear problem with two different iterative schemes. The only difference is that before jumping to N_f, we have both iterations before and after the discontinuous points, i.e., $N = 40$ for (4.168) and $N = 75$ for (4.169). Here we just show near-optimal paths with two different choices of parameters for the nonlinear problem with iterative scheme given by Equation (4.147) in Figure 4.19. Figure 4.19A is the calculation result for $\upsilon = 1.5$ and solution expression (4.168) and Figure 4.19B is the calculation result for $\upsilon = 2.5$ and solution expression (4.169) with corresponding computational costs listed in Table 4.6. In this difficult case, the strategies are able to detect the special parameter $N = 40$ and $N = 75$ and provide as well paths that are less expensive than plain iterations with $N = 100$. Note also that impressively the strategies also provide only one iteration for the maximum N close to 100. We lastly remark that the above Figures show that the hand-fitted parameter 3 might not be adapted to different cases shown above and the study of picking a proper value of this parameter will be a part of future work.

Chapter 5

Application of near-optimal path strategies to nonlinear eigenvalue problem

In this chapter, we study the numerical optimal path in the resolution of Gross-Pitaevskii equation, which is an eigenvalue problem. This is a follow-up work of [38], where the authors introduced a separation of *a posteriori* errors into two parts (iteration and discretisation residuals) but did not provide an optimal strategy.

5.1 Problem description and error analysis

We consider the following energy minimization problem

$$E^* = \min \left\{ E(v) := \frac{1}{2}\int_\Omega (\nabla v)^2 + \frac{1}{2}\int_\Omega V v^2 + \frac{1}{4}\int_\Omega v^4, v \in X, \int_\Omega v^2 = 1 \right\}, \tag{5.1}$$

where Ω is the unit cell $(0, 2\pi)$ of a periodic lattice $\mathcal{R}$ of $\mathbb{R}$ and $X = H^1_\#(\Omega)$ is the Sobolev space defined as

$$H^s_\#(\Omega) := \{v|_\Omega, v \in H^s_{loc}(\mathbb{R}) | v \text{ is } 2\pi\text{-periodic}\},$$

and $X' = H^{-1}_\#(\Omega)$ is the dual space of X. Besides, we assume that $V \in L^\infty(\Omega)$. The well-posedness of the above energy minimization problem can be found in e.g., [13] and it was shown that the above energy minimization problem (5.1) has exactly two solutions with opposite signs: u with $u > 0$ in Ω and $-u$. Besides, the minimizer $u \in X$ of (5.1) is also solution of the Euler equation expressed in the weak form:

$$\begin{cases} \forall v \in X, \int_\Omega \nabla u \cdot \nabla v + \int_\Omega V u v + \int_\Omega u^3 v = \int_\Omega \lambda u v, \\ \int_\Omega u^2 = 1, \end{cases} \tag{5.2}$$

where $\lambda \in \mathbb{R}$ is the Lagrange multiplier associated with the constraint $\|u\|_{L^2} = 1$. Moreover, this weak problem (5.2) is equivalent to the strong problem: find $u \in X$ and $\lambda \in \mathbb{R}$ such that

$$\begin{cases} -\Delta u + Vu + u^3 = \lambda u, \\ \int_\Omega u^2 = 1. \end{cases} \tag{5.3}$$

Besides, here we remark that λ is the smallest eigenvalue (or in the quantum chemistry terminology, the ground state eigenvalue) of the linear operator $A_u \coloneqq -\Delta + V + u^2$ and it is shown that this eigenvalue λ is simple (see, for instance, the Appendix of [13]). The energy functional $E\colon X \to \mathbb{R}$ is twice differentiable and its second order derivative E'' is defined as follows: for any $v, w, z \in X$, we have

$$\langle E''(v)w, z\rangle_{X',X} = \langle A_v w, z\rangle_{X',X} + 2\int_\Omega v^2 wz = \int_\Omega \nabla w \cdot \nabla z + \int_\Omega Vwz + 3\int_\Omega v^2 wz. \tag{5.4}$$

Besides, from [38, Lemma 6.2.1], we have the following coercivity and continuity property: there exists $\gamma_e, \beta_e > 0$ such that for any $v \in X$,

$$\gamma_e \|v\|_{H^1}^2 \leqslant \langle (E''(u) - \lambda)v, v\rangle_{X',X}\,, \tag{5.5}$$

and for any $v, w \in X$,

$$\langle (E''(u) - \lambda)v, w\rangle_{X',X} \leqslant \beta_e \|v\|_{H^1}\ \|w\|_{H^1}, \tag{5.6}$$

where (u, λ) is solution of eigenvalue problem (5.2) with λ being the smallest eigenvalue.

Remark 5.1.1.
If $V \in H^r_\#(\Omega)$ $(r > \frac{1}{2})$, then we have $u \in H^{2+r}_\#(\Omega)$. In what follows, we choose

$$\forall x \in [0, 2\pi], \quad V(x) = 1 + \sum_{k \in \mathbb{N}^*} \frac{cos(kx)}{k^{1.01}}.$$

Here, we pick a more irregular potential function such that the solution is not too regular. In addition, we claim that $V \in L^\infty_\#(\Omega) \cap H^{\frac{1.05}{2}-\epsilon}_\#(\Omega)$ $(\epsilon > 0)$.

In this section, we adapt the setting of Chapter 4.3: we restrict again X to its even part and define discretization space X_N $(N > 0)$ as in (4.9). Then we state the discretized weak problem as follows: For $N \in \mathbb{N}^*$, find $u_N \in X_N$ such that

$$\begin{cases} \forall v_N \in X_N, \quad \int_\Omega \nabla u_N \cdot \nabla v_N + \int_\Omega V u_N v_N + \int_\Omega u_N^3 v_N = \int_\Omega \lambda_N u_N v_N, \\ \int_\Omega u_N^2 = 1. \end{cases} \tag{5.7}$$

The solution u_N of the above problem (5.7) also minimizes the energy functional defined through (5.1) over space X_N,

$$E(u_N) = \min_{v_N \in X_N,\ \|v_N\|_{L^2}=1} E(v_N) =: E_N^*. \tag{5.8}$$

After defining properly the eigenvalue problem, we will next give the *a priori* and *a posteriori* analysis. Fortunately, the *a priori* analysis is established in[13] and the the *a posteriori* analysis can be found in [39, 38]. The only difference in our work is that the periodic cell is $(0, 2\pi)$ but not $(0, 1)$, which only leads to minor amendments. Therefore, we state directly the main results and refer the readers to [39, 38] for a detailed proof. Additionally, in our work we will also show the super convergence property, which is essential for supporting the jump efficiency of the optimal

path problem.

Before stating the *a priori* error estimation, we introduce the following adjoint problem which will be used later: for any $w \in X'$, we consider the weak problem: find $\phi_w \in X$ such that

$$\forall v \in X, \quad \langle (E''(u) - \lambda)\phi_w, v\rangle_{X',X} = \langle w, v\rangle_{X',X}\,, \tag{5.9}$$

where (u, λ) is the solution of problem (5.2). With the coercivity property (5.5) and the continuity property (5.6) at hand, the existence and uniqueness of the solution is proved by applying directly the Lax-Milgram Theorem. In addition, similarly to Estimate (4.23), here we also have the following regularity result: there exists constant $c_a > 0$ independent of w such that

$$\|\phi_w\|_{H^2} \leqslant c_a \|w\|_{L^2}. \tag{5.10}$$

Remark 5.1.2.

Here we remark the difference originating from different lengths of the unit cell: the Gagliardo-Nirenberg-type inequality constant. Following the same derivation in [38], we have the following results:

$$\forall v \in X, \quad \|v\|_{L^\infty}^2 \leqslant \sqrt{\frac{1}{4\pi^2} + 4}\|v\|_{H^1}\, \|v\|_{L^2},$$

and

$$\forall v \in X, \quad \|v\|_{L^\infty} \leqslant \left(\frac{1}{4\pi} + \sqrt{\frac{1}{16\pi^2} + 1}\right)^{\frac{1}{2}} \|v\|_{H^1}.$$

Theorem 5.1.1 (*A Priori* Analysis, see [38] and eventiually [13])**.** *Let $u \in X$ be the weak solution of problem (5.2) with corresponding eigenvalue λ, for $N \in \mathbb{N}^*$, let $u_N \in X_N$ be the weak solution of problem (5.7) with corresponding eigenvalue λ_N, let $E\colon X \to \mathbb{R}$ be the energy functional defined though (5.1), let $V \in H^r_{\#}(\Omega)$ for some $r > \frac{1}{2}$ and let η be defined as*

$$\eta = \frac{\min\{\lambda_2 - \lambda, 2\}}{4(\min\{\lambda_2 - \lambda, 2\} + 2|\lambda| + 2)},$$

where λ_2 is the second smallest eigenvalue of A_u. Then, we have

- *u_N converges strongly to u in $H^1(\Omega)$ when N goes to infinity.*
- *There exists $C^E \in \mathbb{R}_+$ independent of N such that*

$$\eta\|u_N - u\|_{H^1}^2 \leqslant E(u_N) - E(u) \leqslant C^E\|u_N - u\|_{H^1}^2. \tag{5.11}$$

- *There exists $C^\lambda \in \mathbb{R}_+$ independent of N such that*

$$|\lambda - \lambda_N| \leqslant C^\lambda(\|u_N - u\|_{H^1}^2 + \|u_N - u\|_{L^2}). \tag{5.12}$$

- *There exists $C_2^\lambda > 0$ independent of N such that*

$$|\lambda_N - \lambda| \leqslant \frac{C_2^\lambda}{N^{2(r+1)}}. \tag{5.13}$$

- *There exists $N_0 \in \mathbb{N}$ and $C^{H1} \in \mathbb{R}_+$ independent of N such that for any $N \geqslant N_0$, $N \in \mathbb{N}^*$, we have*

$$\|u_N - u\|_{H^1} \leqslant C^{H1} \min_{v_N \in X_N} \|v_N - u\|_{H^1}. \tag{5.14}$$

- *There exists* $N_1 \in \mathbb{N}$ *and* $C^{L2} \in \mathbb{R}_+$ *independent of* N *such that for any* $N \geqslant N_1$, $N \in \mathbb{N}^*$, *we have*

$$\|u_N - u\|_{L^2}^2 \leqslant C^{L2} \|u_N - u\|_{H^1} \min_{\phi_N \in X_N} \|\phi_{u_N - u} - \phi_N\|_{H^1}, \tag{5.15}$$

Corollary 5.1.1.1. *Let* $u \in X$ *be the weak solution of problem* (5.2) *with corresponding eigenvalue* λ *and for* $N \in \mathbb{N}^*$, *let* $u_N \in X_N$ *be the weak solution of problem* (5.7) *with corresponding eigenvalue* λ_N, *then there exits constant* $C_1 > 0$ *such that*

$$\|u - u_N\|_{L^2} \leqslant C_1 N^{-1} \|u - u_N\|_{H^1}, \tag{5.16}$$

Proof. This Corollary is a direct consequence of Estimate (5.15). In fact, for any $N \geqslant N_1$, by picking $\phi_N = \Pi_N \phi_{u_N - u}$, we have

$$\begin{aligned} \|u_N - u\|_{L^2}^2 &\leqslant C^{L2} \|u_N - u\|_{H^1} \|\phi_{u_N - u} - \Pi_N \phi_{u_N - u}\|_{H^1} \\ &\leqslant \frac{C^{L2}}{N} \|u_N - u\|_{H^1} \|\phi_{u_N - u}\|_{H^2} \\ &\leqslant \frac{C^{L2} c_a}{N} \|u_N - u\|_{H^1} \|u_N - u\|_{L^2}, \end{aligned}$$

where we make use of (4.19) and the regularity result of the adjoint problem. From the above derivation we deduce directly (5.16) with $C_1 = C^{L2} c_a$. □

Here we state the super convergence result for which we give a detailed proof following the same idea as that in Theorem 4.1.1.

Theorem 5.1.2. *Let* $u \in X$ *be the weak solution of problem* (5.2) *with corresponding eigenvalue* λ, *for* $N \in \mathbb{N}^*$, *let* $u_N \in X_N$ *be the weak solution of problem* (5.7) *with corresponding eigenvalue* λ_N, *let* $\Pi_N \colon X' \to X_N$ *be the extended* L^2*-orthogonal operator onto space* X_N, *let* $V \in H^s(\mathbb{T})$ *for some* $s > \frac{1}{2}$ *be the potential function in problem* (5.2) *and let us denote by* $r = \min\{2, s\}$. *Then for* N *large enough:*

- *There exits constant* $C_2 > 0$ *such that*

$$\|\Pi_N u - u_N\|_{H_1} \leqslant C_2 N^{-1 - \frac{r}{2}} \|u - u_N\|_{H^1}. \tag{5.17}$$

- *There exits constant* $C_3 > 0$ *such that*

$$\|\Pi_N u - u_N\|_{L^2} \leqslant C_3 N^{-\tau} \|u - u_N\|_{H^1}, \tag{5.18}$$

  where $\tau = \min\{r + 1, \frac{r}{2} + 2\}$.

Specifically, for our chosen potential function V *defined in Remark 5.1.1, we have*

$$\|\Pi_N u - u_N\|_{H_1} \leqslant C_2 N^{-\frac{5.05}{4} + \epsilon} \|u - u_N\|_{H^1} \tag{5.19}$$

and

$$\|\Pi_N u - u_N\|_{L^2} \leqslant C_3 N^{-\frac{3.05}{2} + \epsilon} \|u - u_N\|_{H^1}, \tag{5.20}$$

where $\epsilon > 0$.

Proof. For any $v_N \in X_N$, using Equation (5.7), we deduce that

$$\begin{aligned}
\langle (E''(u)-\lambda)(u_N-u), v_N\rangle_{X',X} &= \langle (A_u-\lambda)(u_N-u), v_N\rangle_{X',X} + 2\int_\Omega u^2(u_N-u)v_N \\
&= \langle (A_u-\lambda)u_N, v_N\rangle_{X',X} + 2\int_\Omega u^2(u_N-u)v_N \\
&= \int_\Omega \nabla u_N\cdot\nabla v_N + \int_\Omega V u_N v_N + \int_\Omega u^2 u_N v_N - \lambda\int_\Omega u_N v_N \\
&\quad + 2\int_\Omega u^2(u_N-u)v_N \\
&= \lambda_N\int_\Omega u_N v_N - \int_\Omega u_N^3 v_N + \int_\Omega u^2 u_N v_N - \lambda\int_\Omega u_N v_N \\
&\quad + 2\int_\Omega u^2(u_N-u)v_N \\
&= (\lambda_N-\lambda)\int_\Omega u_N v_N - \int_\Omega (u_N-u)^2(u_N+2u)v_N.
\end{aligned} \tag{5.21}$$

Then, for any $v_N \in X_N$ we have

$$\begin{aligned}
\langle (E''(u)-\lambda)(u_N-\Pi_N u), v_N\rangle_{X',X} &= \langle (E''(u)-\lambda)(u_N-u), v_N\rangle_{X',X} + \langle (E''(u)-\lambda)(u-\Pi_N u), v_N\rangle_{X',X} \\
&= (\lambda_N-\lambda)\int_\Omega u_N v_N - \int_\Omega (u_N-u)^2(u_N+2u)v_N \\
&\quad + \int_\Omega \nabla(u-\Pi_N u)\cdot\nabla v_N + \int_\Omega V(u-\Pi_N u)v_N \\
&\quad + 3\int_\Omega u^2(u-\Pi_N u)v_N - \lambda\int_\Omega (u-\Pi_N u)v_N \\
&= (\lambda_N-\lambda)\int_\Omega u_N v_N - \int_\Omega (u_N-u)^2(u_N+2u)v_N \\
&\quad + \int_\Omega V(u-\Pi_N u)v_N + 3\int_\Omega u^2(u-\Pi_N u)v_N,
\end{aligned} \tag{5.22}$$

where the last equality comes from the L^2 and H^1 orthogonality. Therefore, by picking $v_N =$

$u_N - \Pi_N u$, it follows from Estimate (5.12) that

$$\begin{aligned}
\langle (E''(u)-\lambda)(u_N-\Pi_N u), u_N-\Pi_N u\rangle_{X',X} =&(\lambda_N-\lambda)\int_\Omega u_N(u_N-\Pi_N u)\\
&-\int_\Omega (u_N-u)^2(u_N+2u)(u_N-\Pi_N u)\\
&+\int_\Omega V(u-\Pi_N u)(u_N-\Pi_N u)\\
&+3\int_\Omega u^2(u-\Pi_N u)(u_N-\Pi_N u)\\
=&(\lambda_N-\lambda)\int_\Omega u_N(u_N-\Pi_N u)\\
&-\int_\Omega (u_N-u)^2(u_N+2u)(u_N-\Pi_N u)\\
&+\int_\Omega (V+u^2+u_N^2+uu_N)(u-u_N)(u_N-\Pi_N u)\\
&+\int_\Omega (V+3u^2)(u_N-\Pi_N u)^2\\
&\leqslant C^\lambda(\|u_N-u\|_{H^1}^2+\|u_N-u\|_{L^2})\|u_N-\Pi_N u\|_{L^2}\\
&+\|V+u^2+u_N^2+uu_N\|_{L^\infty}\|u-u_N\|_{L^2}\ \|u_N-\Pi_N u\|_{L^2}\\
&+\|V+3u^2\|_{L^\infty}\|u_N-\Pi_N u\|_{L^2}^2.
\end{aligned} \tag{5.23}$$

On the other hand, from (5.5) we deduce that

$$\langle (E''(u)-\lambda)(u_N-\Pi_N u), u_N-\Pi_N u\rangle_{X',X} \geqslant \gamma_e\|u_N-\Pi_N u\|_{H^1}^2. \tag{5.24}$$

Combining Estimates (5.23) and (5.24) yields

$$\begin{aligned}
\gamma_e\|u_N-\Pi_N u\|_{H^1}^2 \leqslant& C^\lambda(\|u_N-u\|_{H^1}^2+\|u_N-u\|_{L^2})\ \|u_N-\Pi_N u\|_{L^2}\\
&+\|V+u^2+u_N^2+uu_N\|_{L^\infty}\|u-u_N\|_{L^2}\ \|u_N-\Pi_N u\|_{L^2}\\
&+\|V+3u^2\|_{L^\infty}\|u_N-\Pi_N u\|_{L^2}^2.
\end{aligned} \tag{5.25}$$

Next, we will give an upper bound of $\|u_N-\Pi_N u\|_{L^2}$ using $\|u_N-u\|_{L^2}$. Similar to the proof of Lemma 4.1.1, we will make use of the adjoint equation (5.9). Thus, we have

$$\begin{aligned}
\|u_N-\Pi_N u\|_{L^2}^2 =& \int_\Omega (u_N-\Pi_N u)(u_N-\Pi_N u)\\
=&\langle (E''(u)-\lambda)\phi_{u_N-\Pi_N u}, u_N-\Pi_N u\rangle_{X',X}\\
=&\langle (E''(u)-\lambda)(u_N-\Pi_N u), \phi_{u_N-\Pi_N u}-\Pi_N\phi_{u_N-\Pi_N u}\rangle_{X',X}\\
&+\langle (E''(u)-\lambda)(u_N-\Pi_N u), \Pi_N\phi_{u_N-\Pi_N u}\rangle_{X',X}.
\end{aligned}$$

It follows from Equation (5.22) by picking $v_N = \Pi_N \phi_{u_N - \Pi_N u}$ that

$$\begin{aligned}\|u_N - \Pi_N u\|_{L^2}^2 =& \underbrace{\langle (E''(u) - \lambda)(u_N - \Pi_N u), \phi_{u_N - \Pi_N u} - \Pi_N \phi_{u_N - \Pi_N u} \rangle_{X',X}}_{:=(\mathrm{I})} \\ &+ \underbrace{(\lambda_N - \lambda) \int_\Omega u_N \Pi_N \phi_{u_N - \Pi_N u}}_{:=(\mathrm{II})} \underbrace{- \int_\Omega (u_N - u)^2 (u_N + 2u) \Pi_N \phi_{u_N - \Pi_N u}}_{:=(\mathrm{III})} \\ &+ \underbrace{\int_\Omega (V + 3u^2)(u - \Pi_N u) \Pi_N \phi_{u_N - \Pi_N u}}_{:=(\mathrm{IV})}.\end{aligned} \tag{5.26}$$

For the first term, it follows from the continuity property (5.6) and Estimates (4.19) and (5.10) that

$$\begin{aligned}(\mathrm{I}) \leqslant& \beta_e \|u_N - \Pi_N u\|_{H^1} \, \|\phi_{u_N - \Pi_N u} - \Pi_N \phi_{u_N - \Pi_N u}\|_{H^1} \\ \leqslant& \beta_e N^{-1} \|u_N - \Pi_N u\|_{H^1} \, \|\phi_{u_N - \Pi_N u}\|_{H^2} \\ \leqslant& \underbrace{\beta_e c_a}_{:=C_{\mathrm{I}}} N^{-1} \|u_N - \Pi_N u\|_{H^1} \, \|u_N - \Pi_N u\|_{L^2}.\end{aligned} \tag{5.27}$$

Next, for the second term, applying Estimate (5.13) and making use of the fact that $\|\Pi_N \phi_{u_N - \Pi_N u}\|_{L^2} \leqslant \|\phi_{u_N - \Pi_N u}\|_{L^2} \leqslant \|\phi_{u_N - \Pi_N u}\|_{H^2} \leqslant c_a \|u_N - \Pi_N u\|_{L^2}$ yield that

$$\begin{aligned}(\mathrm{II}) \leqslant& C_2^\lambda N^{-2(s+1)} \|u_N\|_{L^2} \, \|\Pi_N \phi_{u_N - \Pi_N u}\|_{L^2} \\ \leqslant& \underbrace{c_a C_2^\lambda}_{:=C_{\mathrm{II}}} N^{-2(s+1)} \|u_N - \Pi_N u\|_{L^2}.\end{aligned} \tag{5.28}$$

For the third term (III), using similar argument about $\Pi_N \phi_{u_N - \Pi_N u}$ as that for term (II) yields that

$$\begin{aligned}(\mathrm{III}) \leqslant& \left(\frac{1}{4\pi} + \sqrt{\frac{1}{16\pi^2} + 1} \right) \|u_N - u\|_{L^2}^2 \, \|u_N + 2u\|_{H^1} \, \|\Pi_N \phi_{u_N - \Pi_N u}\|_{H^1} \\ \leqslant& \underbrace{\left(\frac{1}{4\pi} + \sqrt{\frac{1}{16\pi^2} + 1} \right) c_a \|u_N + 2u\|_{H^1}}_{:=C_{\mathrm{III}}} \|u_N - u\|_{L^2}^2 \, \|u_N - \Pi_N u\|_{L^2}\end{aligned} \tag{5.29}$$

Now we come to the term (IV). It follows from the orthogonality relation and the Estimate (4.19) that

$$\begin{aligned}(\mathrm{IV}) =& \int_\Omega (u - \Pi_N u) \left[(V + 3u^2) \Pi_N \phi_{u_N - \Pi_N u} \right] - \int_\Omega (u - \Pi_N u) \Pi_N \left[(V + 3u^2) \Pi_N \phi_{u_N - \Pi_N u} \right] \\ \leqslant& \|u - \Pi_N u\|_{L^2} \| (V + 3u^2) \Pi_N \phi_{u_N - \Pi_N u} - \Pi_N \left[(V + 3u^2) \Pi_N \phi_{u_N - \Pi_N u} \right] \|_{L^2}^2 \\ \leqslant& N^{-r} \|u - \Pi_N u\|_{L^2} \, \|(V + 3u^2) \Pi_N \phi_{u_N - \Pi_N u}\|_{H^r} \\ \leqslant& N^{-r} \|V + 3u^2\|_{H^r} \, \|u - \Pi_N u\|_{L^2} \, \|\phi_{u_N - \Pi_N u}\|_{H^2} \\ \leqslant& \underbrace{c_a \|V + 3u^2\|_{H^r}}_{:=C_{\mathrm{IV}}} N^{-r} (\|u - u_N\|_{L^2} + \|u_N - \Pi_N u\|_{L^2}) \|u_N - \Pi_N u\|_{L^2},\end{aligned} \tag{5.30}$$

where in the above derivation we use the regularity analysis that $V \in H^s_{\#}(\Omega)$ $(s > \frac{1}{2})$ implies $u \in H^{2+s}_{\#}(\Omega)$ and thus $V + 3u^2 \in H^s_{\#}(\Omega)$. In summary, combing Estimates (5.27), (5.28), (5.29) and (5.30) yields that

$$\begin{aligned}\|u_N - \Pi_N u\|_{L^2} \leqslant & C_{\mathrm{I}} N^{-1} \|u_N - \Pi_N u\|_{H^1} + C_{\mathrm{II}} N^{-2(s+1)} \|u_N - \Pi_N u\|_{L^2} + C_{\mathrm{III}} \|u_N - u\|^2_{L^2} \\ & + C_{\mathrm{IV}} N^{-r} (\|u - u_N\|_{L^2} + \|u_N - \Pi_N u\|_{L^2}),\end{aligned} \tag{5.31}$$

from which we conclude that

$$\begin{aligned}\left(1 - C_{\mathrm{II}} N^{-2(s+1)} - C_{\mathrm{IV}} N^{-r}\right) \|u_N - \Pi_N u\|_{L^2} \leqslant & C_{\mathrm{I}} N^{-1} \|u_N - \Pi_N u\|_{H^1} + C_{\mathrm{III}} \|u_N - u\|^2_{L^2} \\ & + C_{\mathrm{IV}} N^{-r} \|u - u_N\|_{L^2}.\end{aligned} \tag{5.32}$$

When N is large enough, we have $1 - C_{\mathrm{II}} N^{-2(r+1)} - C_{\mathrm{IV}} N^{-r} \geqslant \frac{1}{2}$ and we deduce that

$$\|u_N - \Pi_N u\|_{L^2} \leqslant 2C_{\mathrm{I}} N^{-1} \|u_N - \Pi_N u\|_{H^1} + 2C_{\mathrm{III}} \|u_N - u\|^2_{L^2} + 2C_{\mathrm{IV}} N^{-r} \|u - u_N\|_{L^2}. \tag{5.33}$$

Combining Estimates (5.25) and (5.33) yields

$$\begin{aligned}\gamma_e \|u_N - \Pi_N u\|^2_{H^1} \leqslant & \left(C^\lambda \|u_N - u\|^2_{H^1} + (C^\lambda + \|V + u^2 + u_N^2 + u u_N\|_{L^\infty}) \|u_N - u\|_{L^2} \right. \\ & \left. + \|V + 3u^2\|_{L^\infty} \|u_N - \Pi_N u\|_{H^1}\right) \|u_N - \Pi_N u\|_{L^2} \\ \leqslant & \left(C^\lambda \|u_N - u\|^2_{H^1} + (C^\lambda + \|V + u^2 + u_N^2 + u u_N\|_{L^\infty}) \|u_N - u\|_{L^2} \right. \\ & \left. + \|V + 3u^2\|_{L^\infty} \|u_N - \Pi_N u\|_{H^1}\right) \left(2C_{\mathrm{I}} N^{-1} \|u_N - \Pi_N u\|_{H^1} \right. \\ & \left. + 2C_{\mathrm{III}} \|u_N - u\|^2_{L^2} + 2C_{\mathrm{IV}} N^{-r} \|u - u_N\|_{L^2}\right) \\ \leqslant & 2C_{\mathrm{I}} N^{-1} \|V + 3u^2\|_{L^\infty} \|u_N - \Pi_N u\|^2_{H^1} + 2\|V + 3u^2\|_{L^\infty} \\ & (C_{\mathrm{III}} \|u_N - u\|_{L^2} + C_{\mathrm{IV}} N^{-r}) \|u - u_N\|_{L^2} \ \|u_N - \Pi_N u\|_{H^1} \\ & + 2C_{\mathrm{I}} N^{-1} \left(C^\lambda \|u_N - u\|^2_{H^1} + (C^\lambda + \|V + u^2 + u_N^2 + u u_N\|_{L^\infty}) \|u_N - u\|_{L^2}\right) \\ & \|u_N - \Pi_N u\|_{H^1} + 2 \left(C^\lambda \|u_N - u\|^2_{H^1} + (C^\lambda + \|V + u^2 + u_N^2 + u u_N\|_{L^\infty}\right) \\ & \|u_N - u\|_{L^2}) \left(C_{\mathrm{III}} \|u_N - u\|_{L^2} + C_{\mathrm{IV}} N^{-r}\right) \|u - u_N\|_{L^2}.\end{aligned} \tag{5.34}$$

Firstly, we claim that for N large enough, we have $\gamma_e - 2C_{\mathrm{I}} N^{-1} \|V + 3u^2\|_{L^\infty} \geqslant \frac{\gamma_e}{2}$. Next, from (5.14) we deduce that

$$\|u_N - u\|_{H^1} \leqslant C^{H1} \|\Pi_N u - u\|_{H^1} \leqslant C^{H1} N^{-1} \|u\|_{H^2}. \tag{5.35}$$

Additionally, combining the above estimates with (5.16) yields

$$\|u_N - u\|_{L^2} \leqslant C_1 C^{H1} N^{-2-r} \|u\|_{H^{2+r}}. \tag{5.36}$$

From the above two estimates, we conclude the following estimate:

- $$\begin{aligned} & 2(C_{\mathrm{III}} \|u_N - u\|_{L^2} + C_{\mathrm{IV}} N^{-r}) \|u - u_N\|_{L^2} \\ \leqslant & 2(C_{\mathrm{III}} C_1 C^{H1} \|u\|_{H^{2+r}} N^{-2-r} + C_{\mathrm{IV}} N^{-r}) C_1 N^{-1} \|u - u_N\|_{H^1} \\ \leqslant & \underbrace{2(C_{\mathrm{III}} C_1 C^{H1} \|u\|_{H^{2+r}} + C_{\mathrm{IV}}) C_1}_{:= C_{\mathrm{V}}} N^{-1-r} \|u - u_N\|_{H^1}. \end{aligned} \tag{5.37}$$

•
$$\begin{aligned} &C^{\lambda}\|u_N-u\|_{H^1}^2+(C^{\lambda}+\|V+u^2+u_N^2+uu_N\|_{L^\infty})\|u_N-u\|_{L^2}\\ \leqslant&\underbrace{\left(C^{\lambda}C^{H1}\|u\|_{H^2}+C_1\left(C^{\lambda}+\|V+u^2+u_N^2+uu_N\|_{L^\infty}\right)\right)}_{:=C_{\mathrm{VI}}}N^{-1}\|u_N-u\|_{H^1}. \end{aligned} \tag{5.38}$$

Inserting the above estimate into (5.34) and supposing that N is large enough yields that

$$\begin{aligned} \frac{\gamma_e}{2}\|u_N-\Pi_N u\|_{H^1}^2\leqslant&\left(\|V+3u^2\|_{L^\infty}C_{\mathrm{V}}N^{-1-r}+2C_{\mathrm{I}}C_{\mathrm{VI}}N^{-2}\right)\|u_N-u\|_{H^1}\ \|u_N-\Pi_N u\|_{H^1}\\ &+C_{\mathrm{V}}C_{\mathrm{VI}}N^{-2-r}\|u_N-u\|_{H^1}^2. \end{aligned} \tag{5.39}$$

We regard the above expression as inequality for a quadratic polynomial with variable $\|u_N-\Pi_N u\|_{H^1}$ and apply the quadratic formula to calculate the roots of the above polynomial. By analyzing the decrease rate of each term in the expression of the positive root, we deduce the quadratic convergence result (5.17). Additionally, by inserting (5.17) into (5.33) and combining Estimates (5.16) and (5.36) we deduce Estimate (5.18), which completes the proof. □

Before stating the *a posteriori* error estimation, all relevant elements need to be defined properly. First of all, we define space of multi-variable $\underline{X}\equiv X\times\mathbb{R}$ and defined the function $\mathrm{F}\colon \underline{X}\to\underline{X}^{'}$ as

$$\forall\underline{u}=(u,\lambda)\in\underline{X},\quad \mathrm{F}(\underline{u})=\begin{pmatrix}-\Delta u+Vu+u^3-\lambda u\\ \int_\Omega u^2-1\end{pmatrix}. \tag{5.40}$$

And for any $\underline{u}=(u,\lambda)\in\underline{X}$ its derivative $D\mathrm{F}_{\underline{u}}\colon \underline{X}\to\underline{X}^{'}$ is defined as

$$\forall(v,\tau)\in\underline{X},\quad D\mathrm{F}_{\underline{u}}(v,\tau)=\begin{pmatrix}-\Delta v+Vv+3u^2v-\tau u-\lambda v\\ 2\int_\Omega uv\end{pmatrix}. \tag{5.41}$$

For $\underline{u}\in\underline{X}$ such that $D\mathrm{F}_{\underline{u}}$ is an isomorphism, we give the following notation:

$$\gamma_{\underline{u}}=\|D\mathrm{F}_{\underline{u}}^{-1}\|_{\underline{X}'\to\underline{X}},\quad L(\alpha)=\sup_{\underline{v}\in B(\underline{u},\alpha)}\|D\mathrm{F}_{\underline{u}-\underline{v}}\|_{\underline{X}\to\underline{X}'},\quad \zeta(\underline{u})=\|\mathrm{F}(\underline{u})\|_{\underline{X}'}, \tag{5.42}$$

where $B(\underline{u},\alpha)$ is the ball in $\underline{X}$ of center $\underline{u}$ and radius α. Apart from the above notation, we also introduce the following linear eigenvalue problem associated with an available numerical approximation $\widetilde{u_N}$ of the discrete solution u_N with $\|\widetilde{u_N}\|_{L^2}=1$: Find $(v_N,\mu_N)\in\underline{X}$ such that

$$\begin{cases}\forall w_N\in X_N,\quad \int_\Omega\nabla v_N\cdot\nabla w_N+\int_\Omega Vv_Nw_N+\int_\Omega\widetilde{u_N}^2v_Nw_N=\int_\Omega\mu_Nv_Nw_N,\\ \int_\Omega v_N^2=1.\end{cases} \tag{5.43}$$

Remark 5.1.3.

The numerical approximation $\widetilde{u_N}$ is obtained, e.g., using iterative scheme. And the approximate eigenvalue $\widetilde{\lambda_N}$ is obtained via

$$\widetilde{\lambda_N}=\int_\Omega(\nabla\widetilde{u_N})^2+\int_\Omega V(\widetilde{u_N})^2+\int_\Omega\widetilde{u_N}^4. \tag{5.44}$$

Therefore, according to the Rayleigh Variational Principle, we always have $\widetilde{\lambda_N}\geqslant\mu_N^{(1)}$ where $\mu_N^{(1)}$ is the smallest eigenvalue of problem (5.43)*. If $\widetilde{\lambda_N}=\mu_N^{(1)}$, the lowest eigenvalue of* (5.43)*, then it's clear that $(\widetilde{u_N},\widetilde{\lambda_N})$ is the solution of weak problem* (5.7)*, i.e., $(\widetilde{u_N},\widetilde{\lambda_N})=(u_N,\lambda_N)$.*

Let us denote all the eigenvalues of problem (5.43) by $\left(\mu_N^{(i)}\right)_{i\geqslant 1}$ arranged in increasing order with corresponding L^2 normalized eigenvector $\left(v_N^{(i)}\right)_{i\geqslant 1}$ and we denote $\widetilde{\delta_N} \coloneqq \mu_N^{(2)} - \mu_N^{(1)}$ the gap between two smallest eigenvalues. As remarked before, for $\widetilde{u_N}$ which is closed enough to the discrete solution u_N, we have $\mu_N^{(1)} < \widetilde{\lambda_N} < \mu_N^{(2)}$ and $\widetilde{u_N}$ close to $v_N^{(1)}$ by picking properly $v_N^{(1)}$ but not $-v_N^{(1)}$. And for $(\widetilde{u_N}, \widetilde{\lambda_N}) \in \underline{X}$, we define the residual of problem (5.3) as

$$R(\widetilde{u_N}, \widetilde{\lambda_N}) := -\Delta\widetilde{u_N} + V\widetilde{u_N} + \widetilde{u_N}^3 - \widetilde{\lambda_N}\widetilde{u_N} \tag{5.45}$$

and we note $\widetilde{R_N} := R(\widetilde{u_N}, \widetilde{\lambda_N})$ for simplicity. Lastly, we define

$$\widetilde{A_N} = \int \widetilde{R_N}\widetilde{u_N}.$$

With all these notations at hand, now we state the error estimation based on the invertibility of the derivative $D\mathrm{F}$ and we refer the readers to [38, Lemma 6.3.1] for a detailed proof.

Theorem 5.1.3 (Coarse *A Posteriori* Error Estimation [38]). *For $N \in \mathbb{N}$, let $\widetilde{u_N} \in X_N$ denote a numerical approximation solution of the discrete solution u_N with $\widetilde{\lambda_N}$ being corresponding eigenvalue obtained via* (5.44)*, let $\left(\mu_N^{(i)}\right)_{i\geqslant 1}$ denote all the eigenvalues of problem* (5.43) *arranged in increasing order with corresponding L^2 normalized eigenvector $\left(v_N^{(i)}\right)_{i\geqslant 1}$ and let $\widetilde{\delta_N} \coloneqq \mu_N^{(2)} - \mu_N^{(1)}$ denote the gap between two smallest eigenvalues, we define the constant*

$$C_N = \min\left\{\frac{N^2}{N^2+1}, \widetilde{\beta_N}\right\} - \frac{\sqrt[4]{\frac{1}{4\pi^2}+4}\|V + 3(\widetilde{u_N})^2 - \widetilde{\lambda_N}\|_{L^2}}{N},$$

where

$$\widetilde{\beta_N} = \frac{\frac{1}{4}\min\left\{\widetilde{\delta_N}, 3\right\}}{\frac{1}{4}\min\left\{\widetilde{\delta_N}, 3\right\} + \|(V + 3(\widetilde{u_N})^2 - \widetilde{\lambda_N})_-\|_{L^\infty} + 1},$$

and $(V + 3(\widetilde{u_N})^2 - \widetilde{\lambda_N})_-$ is the negative part of $V + 3(\widetilde{u_N})^2 - \widetilde{\lambda_N}$. Then for $\widetilde{u_N}$ close enough to u_N such that

$$\|v_N^{(1)} - \widetilde{u_N}\|_{L^\infty} \leqslant \frac{1}{4}\min\left\{1, \frac{\min\left\{\frac{\widetilde{\delta_N}}{2}, \frac{1}{\pi} - \frac{1}{4}\right\}}{2\|v_N^{(1)}\|_{L^\infty}}\right\} \quad \text{and} \quad \widetilde{\lambda_N} - \mu_N^{(1)} \leqslant \min\left\{\frac{1}{2}\widetilde{\delta_N}, \frac{1}{4}\right\},$$

and for N large enough such that $C_N > 0$, $D\mathrm{F}_{(\widetilde{u_N}, \widetilde{\lambda_N})} \colon \widetilde{X} \to \widetilde{X}'$ is an isomorphism. Besides, by denoting $\underline{\widetilde{u_N}} \coloneqq (\widetilde{u_N}, \widetilde{\lambda_N})$ and giving the following computable bounds

$$\begin{aligned}
\gamma_{\underline{\widetilde{u_N}}} &\leqslant \max\left\{\widetilde{I_N}\left(\frac{1}{2} + C_N^{-1}\right), C_N^{-1}\left(1 + \widetilde{I_N}(2C_N^{-1} + 1)\right)\right\} \\
L(t) &\leqslant \left(3\sqrt{\frac{1}{4\pi^2}+4}(t+2) + 4\right)t,
\end{aligned}$$

if $2\gamma_{\underline{\widetilde{u_N}}} L\left(2\gamma_{\underline{\widetilde{u_N}}}\zeta(\underline{\widetilde{u_N}})\right) \leqslant 1$, then there exists a unique solution $\underline{u^*} = (u^*, \lambda^*) \in B\left(\underline{\widetilde{u_N}}, 2\gamma_{\underline{\widetilde{u_N}}}\zeta(\underline{\widetilde{u_N}})\right)$

satisfying $\mathrm{F}(\underline{u^*}) = 0$ *and we have the following a posteriori error estimate*

$$\|\widetilde{u_N} - u^*\|_{H^1} + |\widetilde{\lambda_N} - \lambda^*| \leqslant 2\gamma_{\underline{\widetilde{u_N}}} \|\widetilde{R_N}\|_{H^{-1}}. \tag{5.46}$$

Apart from the above coarse estimate, there is another more accurate error estimate that we state here and refer the readers to [38, Theorem 6.3.1]for a detailed proof.

Theorem 5.1.4 (More Accurate *A Posteriori* Error Estimation [38])**.** *For* $N \in \mathbb{N}$*, let* $\widetilde{u_N} \in X_N$ *denote a numerical approximation solution of the discrete solution* u_N *with* $\widetilde{\lambda_N}$ *being corresponding eigenvalue obtained via* (5.44)*, let* $u \in X$ *be the solution of weak problem* (5.2) *with corresponding eigenvalue* λ*, let* $\left(\mu_N^{(i)}\right)_{i\geqslant 1}$ *denote all the eigenvalues of problem* (5.43) *arranged in increasing order with corresponding* L^2 *normalized eigenvector* $\left(v_N^{(i)}\right)_{i\geqslant 1}$ *and let* $\widetilde{\delta_N} := \mu_N^{(2)} - \mu_N^{(1)}$ *denote the gap between two smallest eigenvalues. Then for any* $\alpha > 1$*, there exist* $N_\alpha \in \mathbb{N}$ *such that for any* $N \geqslant N_\alpha$ *and for* $\widetilde{u_N}$ *close enough to* u_N*, the following inequality holds:*

$$\begin{aligned}
&\frac{1}{2}|\widetilde{\lambda_N} - \lambda| + \sqrt{\frac{1}{2}\left(\frac{1}{2\pi} + \sqrt{\frac{1}{4\pi^2}+4}\right)}\|\widetilde{u_N} - u\|_{L^2}\ \|2\widetilde{u_N} + u\|_{L^\infty} + \left\|\left(V + 3(\widetilde{u_N})^2 - \widetilde{\lambda_N} - 1\right)_-\right\|_{L^\infty} \\
&\left[\sqrt{\frac{1}{4\pi^2}+4}\frac{1}{\widetilde{\beta_N}}\|\widetilde{u_N} - u\|_{H^1}\ \|2\widetilde{u_N} + u\|_{L^2} + \frac{1}{\widetilde{\beta_N}}\left[|\widetilde{\lambda_N} - \mu_N^{(1)}| + \|\widetilde{u_N} - v_N^{(1)}\|_{L^2}\left(M_1\|\widetilde{u_N} - u\|_{H^1}\right.\right.\right. \\
&\left.\left.\left. + 2\|(\widetilde{u_N})^3\|_{L^2} + \|\widetilde{u_N}\|_{L^\infty}^2\ \|\widetilde{u_N} - u\|_{L^2}\right) + |\widetilde{\lambda_N} - \lambda|\right] + \|\widetilde{u_N} - u\|_{L^2}\left(1 + \frac{\|2(\widetilde{u_N})^2\mu_N^{(1)}\|_{H^{-1}}}{\widetilde{\beta_N}}\right) \\
&\left. + \frac{1}{N^2} + \frac{C_r}{\widetilde{\beta_N}}\ \frac{1}{N^{1+r}}\|V + 3(\widetilde{u_N})^2\|_{H^r}\right] \leqslant 1 - \frac{1}{\alpha},
\end{aligned} \tag{5.47}$$

where

$$\begin{aligned}
\widetilde{\beta_N} &= \frac{\mu_N^{(2)} - \mu_N^{(1)}}{\mu_N^{(2)} - \mu_N^{(1)} + \left\|\left(V + (\widetilde{u_N})^2 - \mu_N^{(1)} - 1\right)_-\right\|_{L^\infty}}, \\
M_1 &= 1 + \|V\|_{L^\infty} + \|u\|_{L^\infty}^2 + |\lambda|,
\end{aligned}$$

and $0 \leqslant r \leqslant 1$*,* $C_r = 2^r\left(\frac{1}{4\pi^2}+4\right)^{\frac{1}{4}}$*. Under the condition that the above inequality holds, we have the following a posteriori error estimate:*

$$\begin{aligned}
\|\widetilde{u_N} - u\|_{H^1} \leqslant \alpha\Bigg(&\|R(\widetilde{u_N})\|_{H^{-1}} + \left\|\left(V + 3(\widetilde{u_N})^2 - \widetilde{\lambda_N} - 1\right)_-\right\|_{L^\infty}\left[\frac{1}{\widetilde{\beta_N}}\|\Pi_N R(\widetilde{u_N})\|_{H^{-1}}\right. \\
&\left.\left. + \frac{2}{\widetilde{\beta_N}}|\widetilde{\lambda_N} - \mu_N^{(1)}|\ \|\widetilde{u_N} - v_N^{(1)}\|_{L^2} + \frac{3}{2}\|\widetilde{u_N} - v_N^{(1)}\|_{L^2}^2\left(1 + \frac{\|2(\widetilde{u_N})^2\mu_N^{(1)}\|_{H^{-1}}}{\widetilde{\beta_N}}\right)\right]\right).
\end{aligned} \tag{5.48}$$

Remark 5.1.4.

Here, we remark that compared with the coarse a posteriori error estimator, the second error estimator is not fully computable. In Theorem 5.1.3, the validation condition is $2\gamma_{\underline{\widetilde{u_N}}} L\left(2\gamma_{\underline{\widetilde{u_N}}}\zeta(\underline{\widetilde{u_N}})\right) \leqslant 1$*. Even though we can not compute directly* $\gamma_{\underline{\widetilde{u_N}}}$ *and the Lipschitz function* L*, we still have com-*

putable upper bound of those two terms. As long as the computable upper bound of $2\gamma_{\widetilde{u_N}} L\left(2\gamma_{\widetilde{u_N}}\zeta(\widetilde{u_N})\right)$ *is smaller than* 1*, the validation condition is verified.*

5.2 Iteration scheme and analysis

In this section, we give the details of numerical solution of this problem and the convergence analysis, which is quite different from the source problem in previous chapter. Based on the discretized weak problem (5.7) where $u_N = \sum_{0\leqslant k<N}(\widehat{u_N})_k e_k$ with additional constraint $\|u_N\|_{L^2} = 1$, we select the basis vectors $\{e_i\}_{0\leqslant i<N}$ as test functions v_N and obtain a system of nonlinear eigenvalue equations represented in the following matrix form

$$\begin{cases} \mathbf{A}_N(\mathbf{u}_N)\mathbf{u}_N = \pi\lambda_N\mathbf{u}_N, \\ \mathbf{u}_N^T\mathbf{u}_N = \pi^{-1}, \end{cases} \tag{5.49}$$

where π appearing in the equation originates from the basis vector integration $\int_\Omega e_k^2 = \pi(0 \leqslant k \leqslant N)$. Similar to the nonlinear iterative scheme given by Equation (4.114) in previous chapter, we have the following decomposition of matrix:

$$\mathbf{A}_N(\mathbf{u}_N) = \mathbf{U}_N + \mathbf{D}_N + \mathbf{L}_N + \mathbf{S}_N(\mathbf{u}_N),$$

and we still make use of the Gauss-Seidel-Relaxation (GSR) iterative scheme. By setting an initial guess $(\mathbf{u}_N^0, \lambda_N^0)$, in each iteration the new eigenpair $(\mathbf{u}_N^{k+1}, \lambda_N^{k+1})$ is calculated from $(\mathbf{u}_N^k, \lambda_N^k)$ via the following process:

$$\begin{cases} \left(\mathbf{u}_N^{k+1}\right)^* = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}_N(\mathbf{u}_N^k)\right)\mathbf{u}_N^k + \pi(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\lambda_N^k\mathbf{u}_N^k, \\ \mathbf{u}_N^{k+1} = \frac{\left(\mathbf{u}_N^{k+1}\right)^*}{\pi^{\frac{1}{2}}\|\left(\mathbf{u}_N^{k+1}\right)^*\|_{L^2}}, \\ \lambda_N^{k+1} = (\mathbf{u}_N^{k+1})^T\mathbf{A}_N(\mathbf{u}_N^{k+1})\mathbf{u}_N^{k+1}, \end{cases} \tag{5.50}$$

with ω being the relaxation factor. Here we remark that in the eigenvalue problem there is no oscillation of the numerical solutions as indicated in Remark 4.2.2. Therefore, unlike iterative scheme given by Equation (4.114), we don't add the damping term in the iterative scheme given by Equation (5.50). From (5.50), we know that

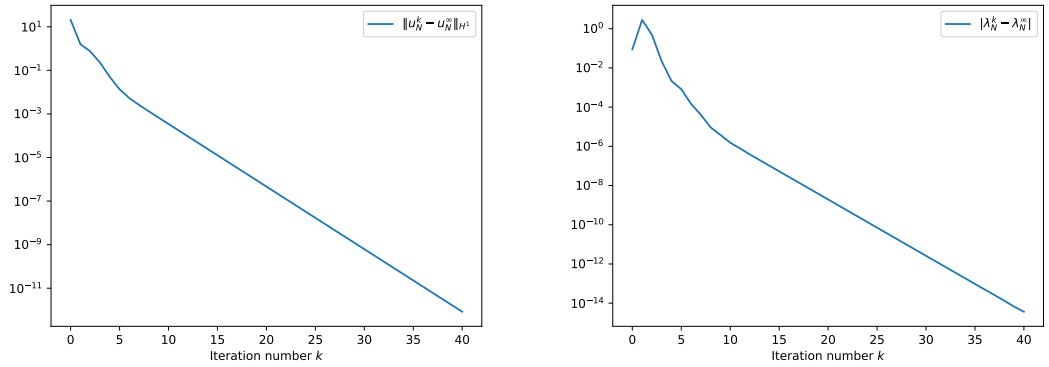
$$\left(\mathbf{u}_N^{k+1}\right)^* = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}_N(\mathbf{u}_N^k) - \pi\lambda_N^k\mathbf{I}\right)\mathbf{u}_N^k.$$

Therefore, by denoting $\mathbf{P}_N(\mathbf{u}_N^k, \lambda_N^k) = -(\mathbf{D}_N + \omega\mathbf{L}_N)^{-1}\left((1-\omega)\mathbf{L}_N + \mathbf{U}_N + \mathbf{S}_N(\mathbf{u}_N^k) - \pi\lambda_N^k\mathbf{I}\right)$, the above GSR scheme has the following equivalent expression:

$$\begin{cases} \left(\mathbf{u}_N^{k+1}\right)^* = \mathbf{P}_N(\mathbf{u}_N^k, \lambda_N^k)\mathbf{u}_N^k, \\ \mathbf{u}_N^{k+1} = \frac{\left(\mathbf{u}_N^{k+1}\right)^*}{\pi^{\frac{1}{2}}\|\left(\mathbf{u}_N^{k+1}\right)^*\|_{L^2}}, \\ \lambda_N^{k+1} = (\mathbf{u}_N^{k+1})^T\mathbf{A}_N(\mathbf{u}_N^{k+1})\mathbf{u}_N^{k+1}. \end{cases} \tag{5.51}$$

The main difficulty for the analysis of the iterative scheme (5.51) relies on the update of matrix $\mathbf{P}_N(\mathbf{u}_N^k, \lambda_N^k)$ in each iteration. From Estimate (5.12), we know that the convergence rate of eigenvalue is faster than the eigenvector. When λ_N^k is close enough to λ_N^∞, by assuming that the influence of the nonlinear term $\mathbf{S}_N(\mathbf{u}_N^k)$ is small compared to the contribution from other terms, we can deduce that the decrease rate of the iteration error is bounded by the matrix norm of

$\pi^{-\frac{1}{2}}\|\mathbf{P}_N(\mathbf{u}_N^\infty,\lambda_N^\infty)\mathbf{u}_N^\infty\|_{L^2}^{-1}\mathbf{P}_N(\mathbf{u}_N^\infty,\lambda_N^\infty)$. Additionally, we perform the following numerical test: For $N = 100$, given $u_N^0 = (0.01,\cdots,0.01)$ as the initial guess, we perform 40 GSR iterations according to Scheme (5.51) with $\omega = 0.2$ and plot the variation of eigenvector errors $\|u_N^k - u_N^\infty\|_{H^1}$ and eigenvalue errors $|\lambda_N^k - \lambda_N^\infty|$ in Figure 5.1. From Figure 5.1, firstly we observe the convergence for both the eigenvalue and eigenvector and the faster convergence of eigenvalue. In addition, after the first 5 iterations, the variation of the eigenvalue λ_N^k becomes small, i.e., $\lambda_N^k \approx \lambda_N^\infty$, and we observe that the corresponding iteration error $\|u_N^k - u_N^\infty\|_{H^1}$ converges to 0 with a near-constant decrease rate, which is similar to the convergence result of the source problems.



(A) The variation of $\|u_N^k - u_N^\infty\|_{H^1}$ for $N = 100$ as a function of iteration number k.

(B) The variation of $|\lambda_N^k - \lambda_N^\infty|$ for $N = 100$ as a function of iteration number k.

Figure 5.1: The convergence of nonlinear GSR eigen scheme .

5.3 Strategies and numerical test

In this section, we will explore the optimal path problem for the solution of eigenvalue problem (5.2). Thanks to the work for the solution of the linear and nonlinear source problems, now we don't need to apply probabilistic method to explore the optimal path and then to summarize the characteristics of the optimal path. In this section, we will skip this time-consuming process and focus on checking several key properties of the eigenvalue problem such that we can directly apply the same strategies to give near-optimal paths. First at all, we decompose the total residual into two parts: the discretization residual and the iteration residual. Recall the definition of the residual in (5.45) and the iterative scheme given by Equation (5.50), for numerical solution (u_N^k,λ_N^k) and the intermediate solution $(u_N^k)^*$, we define the discretization residual as

$$R_{\text{disc}}(u_N^k,\lambda_N^k) := A_1\left(u_N^k - \frac{u_N^{k-1}}{\|(u_N^k)^*\|_{L^2}}\right) + \frac{1}{\|(u_N^k)^*\|_{L^2}}R(u_N^{k-1},\lambda_N^{k-1}), \tag{5.52}$$

and the iteration residual as

$$\begin{aligned} R_{\text{iter}}(u_N^k, \lambda_N^k) :=& R(u_N^k, \lambda_N^k) - R_{\text{disc}}(u_N^k, \lambda_N^k) \\ =& A_2 \left(u_N^k - \frac{u_N^{k-1}}{\|(u_N^k)^*\|_{L^2}} \right) + (u_N^k)^3 - \frac{1}{\|(u_N^k)^*\|_{L^2}} (u_N^{k-1})^3 \\ & - \left(\lambda_N^k - u_N^k - \frac{1}{\|(u_N^k)^*\|_{L^2}} \lambda_N^{k-1} - u_N^{k-1} \right), \end{aligned} \tag{5.53}$$

where A_1 and A_2 are defined in Corollary 4.1.1.1.

Next, the first key property to check is the jump efficiency. In fact, the iterative scheme given by Equation (5.51) also maintains the jump efficiency as for the source problems. Here we give one numerical example to illustrate the jump efficiency in this case: Firstly, we set the initial guess for $N = 3$ as $\mathbf{u}_N^0 = \{0.01, 0.01, 0.01, 0.01\}$ and here we remind the reader that initial guess can not be zero, otherwise the numerical solution will always be zero. Then, we perform 10 iterations for $N = 3$ and we calculate the ratio $\frac{\|\mathbf{u}_N^{k+1} - \mathbf{u}_N^\infty\|_{H^1}}{\|\mathbf{u}_N^k - \mathbf{u}_N^\infty\|_{H^1}}$ for those 10 iterations. Those values vary between 0.23 and 1.27 and the value 1.27 corresponds to the ratio for the first iteration. Then those rest ratios are all smaller that 1 and they vary from 0.28 to 0.515. With the increase of iteration numbers, the ratio is closer and closer to 0.515. Then we switch from $N = 3$ to $N' = 12$ and perform 4 iterations. We calculate the ratio $\frac{\|\mathbf{u}_{N'}^{k+1} - \mathbf{u}_{N'}^\infty\|_{H^1}}{\|\mathbf{u}_{N'}^k - \mathbf{u}_{N'}^\infty\|_{H^1}}$ for these 4 iterations. The ratios are about $0.072, 0.245, 0.298, 0.425$. From this simple test, we observe the same jump efficiency as in the linear and nonlinear source problems.

The following key property to be checked is the predictability of N_f. Recalling Scheme (5.50), then truncating the residual $R(u_N^k, \lambda_N^k)$ at $N'(N' > N)$ yields that

$$\begin{aligned} \Pi_{N'} R(u_N^k, \lambda_N^k) =& \Pi_{N'} A_1 u_N^k + \Pi_{N'} A_2 u_N^k + \Pi_{N'} (u_N^k)^3 - \lambda_N^k u_N^k \\ =& \Pi_{N'} A_1 u_N^k - \Pi_{N'} A_1 (u_{N'}^1)^* \\ =& \Pi_{N'} A_1 (u_N^k - (u_{N'}^1)^*), \end{aligned}$$

where $(u_{N'}^1)*$ is the intermediate solution obtained via one GSR iteration after the jump from N to N'. In addition, we have

$$\Pi_{N'}^\perp R(u_N^k, \lambda_N^k) = \Pi_{N'}^\perp A(u_N^k - u_{N'}^\infty) + \Pi_{N'}^\perp \left((u_N^k)^3 - (u_{N'}^\infty)^3 \right) + R(u_{N'}^\infty, \lambda_{N'}^\infty).$$

Therefore, if we assume that the term $\Pi_{N'} A(u_N^k - u_N^\infty) + \Pi_{N'} \left((u_N^k)^3 - (u_{N'}^\infty)^3 \right)$ is dominant in $A(u_N^k - u_N^\infty) + \left((u_N^k)^3 - (u_{N'}^\infty)^3 \right)$, then the term $\Pi_{N'}^\perp R(u_N^k)$ can represent $R_{N'}^\infty$. Here, we do one test: For $N = 3$, we calculate the residual R_N^∞ and then for $N < N' \leqslant 100$, we compare $\|\Pi_{N'}^\perp R_N^\infty\|_{H^{-1}}$ and $\|R_{N'}^\infty\|_{H^{-1}}$. The result shows that for each $N < N' \leqslant 100$, we always have $\|\Pi_{N'}^\perp R_N^\infty\|_{H^{-1}} > \|R_{N'}^\infty\|_{H^{-1}}$. This result confirms that in the eigenvalue problem, we can also predict N_f such that $R_{N_f}^\infty \leqslant \varepsilon_g$.

According to the above tests, we can already conclude that it's feasible to apply Strategy I (Algorithm 3) for the solution of eigenvalue problem. For applying the Strategy II, we still need to check the behavior of the iteration residual after the jump. The first observation is that the decrease of the iteration errors for calculations with a fixed N is nearly a constant with respect to N. Then we do the same test as that for the source problems: For $N = 5$, we take u_N^∞ as input to calculate output $u_{N'}^1$ for different $N < N' \leqslant 100$ and the corresponding iteration residual. We plot the result in Figure 5.2. From Figure 5.2, we observe that the iteration residual increases for N' close to initial input N and then it's nearly a constant for other choice

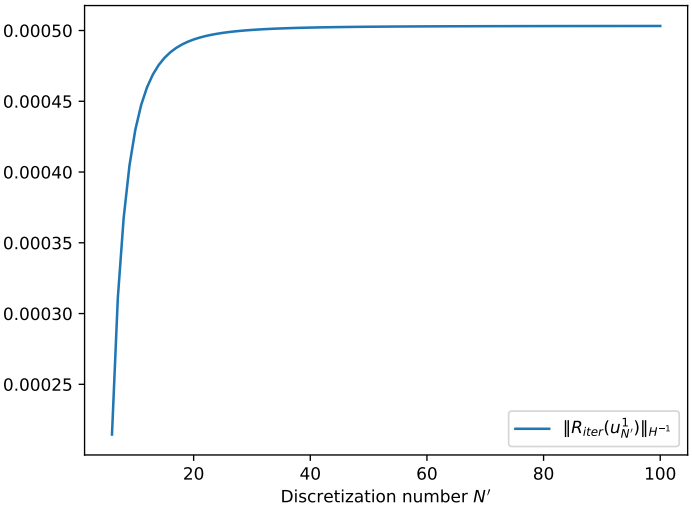


Figure 5.2: Iteration residual of $u^1_{N'}$ for different discretization number N'.

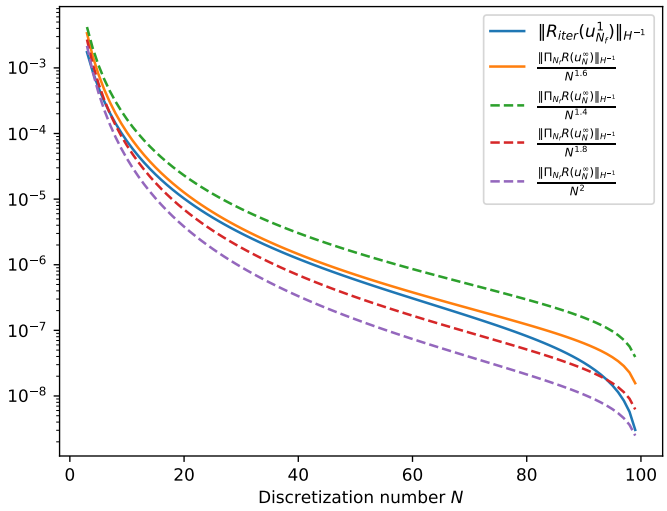


Figure 5.3: Iteration residual of u^1_{100} for different discretization number N before the jump.

of output discretization number N', which is the same as that for source problems. Next, We fix $N_f = 100$, take u^∞_N as input for $3 \leqslant N \leqslant N_f - 1$, jump from N to N_f, perform one GSR iteration to obtain output $u^1_{N_f}$ and calculate the iteration residual $\|R_{\text{iter}}(u^1_{N_f})\|_{H^{-1}}$. Then we plot the variation of $\|R_{\text{iter}}(u^1_{N_f})\|_{H^{-1}}$ and $\frac{\|\Pi_{N_f} R(u^\infty_N)\|_{H^{-1}}}{N^{1.6}}$ in Figure 5.3. Here, the parameter connecting the discretization residual and the iteration residual is not N^2 but $N^{1.6}$, which is the

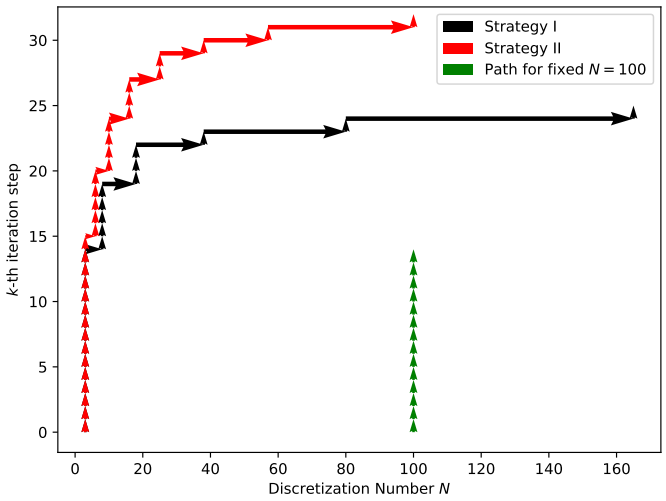


Figure 5.4: Nearly optimal strategies for the eigenvalue problem

same as that for the nonlinear source case, which indicates that this difference might originate from the addition of the nonlinear term. From this figure, we conclude that it is reasonable to add this criteria into our strategy.

Path	Fixed $N = 100$	Strategy I	Strategy II
Cost	14,567,028	5,227,524	1,360,896

Table 5.1: Computational costs of different paths for the eigenvalue problem.

By summarizing all above arguments, we conclude that we can also apply Strategy II for the numerical solution of our eigenvalue optimal path problem. By setting the goal residual as $\varepsilon_g = 0.5\varepsilon_{99} + 0.5\varepsilon_{100}$ and transferring it into the residual requirement, we compare three paths: path for fixed $N = 100$ and paths generated by Strategy I and II respectively. We plot the result in Figure 5.4 and list corresponding computational costs in Table 5.1. From the figure, we see that both the strategies can finish the path with only 1 iteration for the biggest discretization number N. Unfortunately, Strategy I can not predict well the last discretization number N_f, which increase the computational cost but Strategy II still behaves well in this case.

Bibliography

[1] I. Althöfer and K.-U. Koschnick. "On the convergence of "Threshold Accepting"". In: *Applied Mathematics and Optimization* 24.1 (1991), pp. 183–195.

[2] *Annual Report 2021 of the Swiss National Supercomputing Centre*. `https://www.cscs.ch/fileadmin/user_upload/contents_publications/annual_reports/AR2021_Final_WEB.pdf`.

[3] I. Babuška and W. C. Rheinboldt. "A-posteriori error estimates for the finite element method". In: *International journal for numerical methods in engineering* 12.10 (1978), pp. 1597–1615.

[4] W. Bangerth and R. Rannacher. *Adaptive finite element methods for differential equations*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, 2003, pp. viii+207.

[5] M. P. Barnett. "Mechanized Molecular Calculations—The POLYATOM System". In: *Reviews of Modern Physics* 35.3 (1963), p. 571.

[6] K.-J. Bathe. "Finite element method". In: *Wiley encyclopedia of computer science and engineering* (2007), pp. 1–12.

[7] M. J. Berger and P. Colella. "Local adaptive mesh refinement for shock hydrodynamics". In: *Journal of computational Physics* 82.1 (1989), pp. 64–84.

[8] S. F. Boys. "Electronic wave functions-I. A general method of calculation for the stationary states of any molecular system". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 200.1063 (1950), pp. 542–554.

[9] S. C. Brenner. *The mathematical theory of finite element methods*. Springer, 2008.

[10] H. Brezis. *Functional analysis, Sobolev spaces and partial differential equations*. Springer-Verlag New York, 2011.

[11] H. Brézis. *Functional analysis, Sobolev spaces and partial differential equations*. Vol. 2. 3. Springer, 2011.

[12] G. Caloz and J. Rappaz. "Numerical analysis for nonlinear and bifurcation problems". In: *Handbook of numerical analysis, Vol. V*. Handbook of Numerical Analysis, Vol. V. North-Holland, Amsterdam, 1997, pp. 487–637.

[13] E. Cancès, R. Chakir, and Y. Maday. "Numerical analysis of nonlinear eigenvalue problems". In: *Journal of Scientific Computing* 45.1-3 (2010), pp. 90–117.

[14] E. Cancès, R. Chakir, and Y. Maday. "Numerical analysis of the planewave discretization of some orbital-free and Kohn-Sham models". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46.2 (2012), pp. 341–388.

[15] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. "Computational quantum chemistry: a primer". In: *Handbook of numerical analysis, Vol. X*. Handb. Numer. Anal., X. North-Holland, Amsterdam, 2003, pp. 3–270.

[16] E. Cancès, G. Dusson, Y. Maday, B. Stamm, and M. Vohralík. "A perturbation-method-based *a posteriori* estimator for the planewave discretization of nonlinear Schrödinger equations". en. In: *Comptes Rendus. Mathématique* 352.11 (2014), pp. 941–946.

[17] E. Cances and C. Le Bris. "On the convergence of SCF algorithms for the Hartree-Fock equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 34.4 (2000), pp. 749–774.

[18] C. Canuto, M. Y. Hussaini, A. Quarteroni, and T. A. Zang. *Spectral methods: fundamentals in single domains*. Springer Science & Business Media, 2007.

[19] J. Cea. "Approximation variationnelle des problèmes aux limites". fr. In: *Annales de l'Institut Fourier* 14.2 (1964), pp. 345–444.

[20] F. Chatelin. *Spectral approximation of linear operators*. SIAM, 2011.

[21] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou. "Adaptive Finite Element Approximations for Kohn–Sham Models". In: *Multiscale Modeling & Simulation* 12.4 (2014), pp. 1828–1869. eprint: `https://doi.org/10.1137/130916096`.

[22] H. Chen, X. Gong, L. He, and A. Zhou. "Adaptive Finite Element Approximations for a Class of Nonlinear Eigenvalue Problems in Quantum Physics". In: *Advances in Applied Mathematics and Mechanics* 3.4 (2011), 493–518.

[23] H. Chen, X. Gong, and A. Zhou. "Numerical approximations of a nonlinear eigenvalue problem and applications to a density functional model". In: *Mathematical methods in the applied sciences* 33.14 (2010), pp. 1723–1742.

[24] *CINECA 2020-2021 HPC Annual Report*. `https://www.cineca.it/sites/default/files/2021-11/REPORT_HPC_20202021.pdf`.

[25] J. Čížek. "On the correlation problem in atomic and molecular systems. Calculation of wavefunction components in Ursell-type expansion using quantum-field theoretical methods". In: *The Journal of Chemical Physics* 45.11 (1966), pp. 4256–4266.

[26] E Clementi and D. Davis. "Electronic structure of large molecular systems". In: *Journal of Computational Physics* 1.2 (1966), pp. 223–244.

[27] F. Coester. "Bound states of a many-particle system". In: *Nuclear Physics* 7 (1958), pp. 421–424.

[28] F. Coester and H. Kümmel. "Short-range correlations in nuclear wave functions". In: *Nuclear Physics* 17 (1960), pp. 477–485.

[29] M. A. Csirik and A. Laestadius. "Coupled-Cluster theory revisited. Part I: Discretization". In: *ESAIM: Mathematical Modelling and Numerical Analysis* (2022).

[30] M. A. Csirik and A. Laestadius. "Coupled-Cluster theory revisited. Part II: Analysis of the single-reference Coupled-Cluster equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* (2022).

[31] X. Dai, L. He, and A. Zhou. "Convergence and quasi-optimal complexity of adaptive finite element computations for multiple eigenvalues". In: *IMA Journal of Numerical Analysis* 35.4 (2015), pp. 1934–1977.

[32] X. Dai, J. Xu, and A. Zhou. "Convergence and optimal complexity of adaptive finite element eigenvalue computations". In: *Numerische Mathematik* 110.3 (2008), pp. 313–355.

[33] P. Deglmann, A. Schäfer, and C. Lennartz. "Application of quantum calculations in the chemical industry—An overview". In: *International Journal of Quantum Chemistry* 115.3 (2015), pp. 107–136.

[34] G. Dhatt, E. Lefrançois, and G. Touzot. *Finite element method*. John Wiley & Sons, 2012.

[35] J. M. Dieterich and E. A. Carter. "Opinion: Quantum solutions for a sustainable energy future". In: *Nature Reviews Chemistry* 1.4 (2017), pp. 1–7.

[36] P. A. M. Dirac and R. H. Fowler. "On the theory of quantum mechanics". In: *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character* 112.762 (1926), pp. 661–677. eprint: `https://royalsocietypublishing.org/doi/pdf/10.1098/rspa.1926.0133`.

[37] G. Dueck and T. Scheuer. "Threshold accepting: A general purpose optimization algorithm appearing superior to simulated annealing". In: *Journal of Computational Physics* 90.1 (1990), pp. 161–175.

[38] G. Dusson. "Estimation d'erreur pour des problèmes aux valeurs propres linéaires et non-linéaires issus du calcul de structure électronique". Thèse de doctorat dirigée par Maday, Yvon et Piquemal, Jean-Philip Mathématiques appliquées Paris 6 2017. PhD thesis. 2017.

[39] G. Dusson and Y. Maday. "A Posteriori Analysis of a Non-Linear Gross-Pitaevskii type Eigenvalue Problem". In: *IMA Journal of Numerical Analysis* 37 (Aug. 2013).

[40] L. C. Evans. *Partial differential equations*. Vol. 19. American Mathematical Society, 2022.

[41] *Facts and Figures 2020: Forchungszentrum Jülich*. `https://www.fz-juelich.de/en/press/factsandfigures`.

[42] F. M. Faulstich, A. Laestadius, O. Legeza, R. Schneider, and S. Kvaal. "Analysis of the tailored coupled-cluster method in quantum chemistry". In: *SIAM Journal on Numerical Analysis* 57.6 (2019), pp. 2579–2607.

[43] F. M. Faulstich and M. Oster. "Coupled cluster theory: Towards an algebraic geometry formulation". In: *arXiv preprint arXiv:2211.10389* (2022).

[44] V. Fock. "Näherungsmethode zur Lösung des quantenmechanischen Mehrkörperproblems". In: *Zeitschrift für Physik* 61.1 (1930), pp. 126–148.

[45] G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Pure and Applied Mathematics: A Wiley Series of Texts, Monographs and Tracts. Wiley, 2013.

[46] M. J. Frisch et al. *Gaussian16 Revision C.01*. Gaussian Inc. Wallingford CT. 2016.

[47] G. Hall. "The molecular orbital theory of chemical valency VIII. A method of calculating ionization potentials". In: *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* 205.1083 (1951), pp. 541–552.

[48] D. R. Hartree. "The wave mechanics of an atom with a non-Coulomb central field. Part I. Theory and methods". In: *Mathematical Proceedings of the Cambridge Philosophical Society* 24.1 (1928), pp. 89–110.

[49] M. Hassan, Y. Maday, and Y. Wang. "Analysis of the single reference coupled cluster method for electronic structure calculations: the full-coupled cluster equations". In: *Numerische Mathematik* 155.1 (2023), pp. 121–173.

[50] W. Heisenberg. "Mehrkörperproblem und Resonanz in der Quantenmechanik". In: *Zeitschrift für Physik* 38.6 (1926), pp. 411–426.

[51] W. Heitler and F. London. "Wechselwirkung neutraler Atome und homöopolare Bindung nach der Quantenmechanik". In: *Zeitschrift für Physik* 44.6 (1927), pp. 455–472.

[52] T. Helgaker, P. Jorgensen, and J. Olsen. *Molecular electronic-structure theory*. John Wiley & Sons, 2014.

[53] A. Hillisch, N. Heinrich, and H. Wild. "Computational chemistry in the pharmaceutical industry: from childhood to adolescence". In: *ChemMedChem* 10.12 (2015), pp. 1958–1962.

[54] G. Hu, H. Xie, and F. Xu. "A multilevel correction adaptive finite element method for Kohn–Sham equation". In: *Journal of Computational Physics* 355 (2018), pp. 436–449.

[55] W. Hunziker and I. M. Sigal. "The quantum N-body problem". In: *Journal of Mathematical Physics* 41.6 (2000), pp. 3448–3510.

[56] E. A. Hylleraas. "Neue berechnung der energie des heliums im grundzustande, sowie des tiefsten terms von ortho-helium". In: *Zeitschrift für Physik* 54.5 (1929), pp. 347–366.

[57] E. A. Hylleraas. "Über den grundzustand des heliumatoms". In: *Zeitschrift für Physik* 48.7 (1928), pp. 469–494.

[58] S. Kirkpatrick, C. D. Gelatt Jr, and M. P. Vecchi. "Optimization by simulated annealing". In: *science* 220.4598 (1983), pp. 671–680.

[59] K. Kowalski and K. Jankowski. "Towards Complete Solutions to Systems of Nonlinear Equations of Many-Electron Theories". In: *Physical Review Letters* 81 (6 1998), pp. 1195–1198.

[60] A. Laestadius and F. M. Faulstich. "The coupled-cluster formalism–a mathematical perspective". In: *Molecular Physics* 117.17 (2019), pp. 2362–2373.

[61] A. Laestadius and S. Kvaal. "Analysis of the extended coupled-cluster method in quantum chemistry". In: *SIAM Journal on Numerical Analysis* 56.2 (2018), pp. 660–683.

[62] T. J. Lee and G. E. Scuseria. "Achieving chemical accuracy with coupled-cluster theory". In: *Quantum Mechanical Electronic Structure Calculations with Chemical Accuracy*. Springer, 1995, pp. 47–108.

[63] D. Li, G. Ströhmer, and L. Wang. "Symmetry of integral equations on bounded domains". In: *Proceedings of the American Mathematical Society* 137.11 (2009), pp. 3695–3702.

[64] K. Lipnikov, G. Manzini, and M. Shashkov. "Mimetic finite difference method". In: *Journal of Computational Physics* 257 (2014), pp. 1163–1227.

[65] S. L. Lovelock et al. "The road to fully programmable protein catalysis". In: *Nature* 606.7912 (2022), pp. 49–58.

[66] D. G. Luenberger. *Optimization by vector space methods*. John Wiley & Sons, Inc., New York-London-Sydney, 1969, pp. xvii+326.

[67] Y. Maday and G. Turinici. "Error bars and quadratically convergent methods for the numerical simulation of the Hartree-Fock equations". In: *Numerische Mathematik* 94.4 (2003), pp. 739–770.

[68] C. J. Manly, S. Louise-May, and J. D. Hammer. "The impact of informatics and computational chemistry on synthesis and screening". In: *Drug discovery today* 6.21 (2001), pp. 1101–1110.

[69] P. Moczo, J. Kristek, and L. Halada. "The finite-difference method for seismologists". In: *An Introduction* 161 (2004).

[70] M. Nooijen*, K. Shamasundar, and D. Mukherjee. "Reflections on size-extensivity, size-consistency and generalized extensivity in many-body theory". In: *Molecular Physics* 103.15-16 (2005), pp. 2277–2298.

[71] J. T. Oden, L. Demkowicz, W. Rachowicz, and T. A. Westermann. "Toward a universal hp adaptive finite element strategy, Part 2. A posteriori error estimation". In: *Computer methods in applied mechanics and engineering* 77.1-2 (1989), pp. 113–180.

[72] M Ortiz and J. Quigley Iv. "Adaptive mesh refinement in strain localization problems". In: *Computer Methods in Applied Mechanics and Engineering* 90.1-3 (1991), pp. 781–804.

[73] M. N. Özişik, H. R. Orlande, M. J. Colaço, and R. M. Cotta. *Finite difference methods in heat transfer*. CRC press, 2017.

[74] J Paldus, J Čížek, and I Shavitt. "Correlation Problems in Atomic and Molecular Systems. IV. Extended Coupled-Pair Many-Electron Theory and Its Application to the BH_3 Molecule". In: *Physical Review A* 5.1 (1972), p. 50.

[75] J. Paldus, P. Piecuch, L. Pylypow, and B. Jeziorski. "Application of Hilbert-space coupled-cluster theory to simple $(H_2)_2$ model systems: Planar models". In: *Physical Review A* 47 (4 1993), pp. 2738–2782.

[76] N. Perrone and R. Kao. "A general finite difference method for arbitrary meshes". In: *Computers & Structures* 5.1 (1975), pp. 45–57.

[77] P. Piecuch and K. Kowalski. "In Search of the Relationship between Multiple Solutions Characterizing Coupled-Cluster Theories". In: *Computational Chemistry: Reviews of Current Trends*. 2000, pp. 1–104.

[78] P. Piecuch, S. Zarrabian, J. Paldus, and J. Čížek. "Coupled-cluster approaches with an approximate account of triexcitations and the optimized-inner-projection technique. II. Coupled-cluster results for cyclic-polyene model systems". In: *Physical Review B* 42 (6 1990), pp. 3351–3379.

[79] K. Raghavachari, G. W. Trucks, J. A. Pople, and M. Head-Gordon. "A fifth-order perturbation comparison of electron correlation theories". In: *Chemical Physics Letters* 157.6 (1989), pp. 479–483.

[80] J. N. Reddy. *Introduction to the finite element method*. McGraw-Hill Education, 2019.

[81] T. Rohwedder. "An analysis for some methods and algorithms of Quantum Chemistry". PhD thesis. Technische Universität Berlin, 2010.

[82] T. Rohwedder. "The continuous coupled cluster formulation for the electronic Schrödinger equation". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 47.2 (2013), pp. 421–447.

[83] T. Rohwedder and R. Schneider. "Error estimates for the coupled cluster method". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 47.6 (2013), pp. 1553–1582.

[84] C. Roothaan. "Self-consistent field theory for open shells of electronic systems". In: *Reviews of modern physics* 32.2 (1960), p. 179.

[85] C. Roothaan. "New developments in molecular orbital theory". In: *Reviews of modern physics* 23.2 (1951), p. 69.

[86] A. Savin. "Is size-consistency possible with density functional approximations?" In: *Chemical Physics* 356.1-3 (2009), pp. 91–97.

[87] R. Schneider. "Analysis of the projected coupled cluster method in electronic structure calculation". In: *Numerische Mathematik* 113.3 (2009), pp. 433–471.

[88] K. Segeth. "A review of some a posteriori error estimates for adaptive finite element methods". In: *Mathematics and Computers in Simulation* 80.8 (2010), pp. 1589–1600.

[89] O. Sinanoğlu. "Many-Electron Theory of Atoms and Molecules. I. Shells, Electron Pairs vs Many-Electron Correlations". In: *The Journal of Chemical Physics* 36.3 (1962), pp. 706–717.

[90] J. C. Slater. "Atomic shielding constants". In: *Physical Review* 36.1 (1930), p. 57.

[91] J. C. Slater. "Note on Hartree's method". In: *Physical Review* 35.2 (1930), p. 210.

[92] H. M. Soner, W. Bangerth, R. Rannacher, H. Foellmer, and L. Rogers. *Adaptive finite element methods for differential equations*. Springer Science & Business Media, 2003.

[93] A. Szabo and N. S. Ostlund. *Modern quantum chemistry: introduction to advanced electronic structure theory*. Courier Corporation, 2012.

[94] *The Nobel Prize in Chemistry 1998 Press Release*. `https://www.nobelprize.org/prizes/chemistry/1998/press-release/`.

[95] *The Nobel Prize in Chemistry 2013 Press Release*. `https://www.nobelprize.org/prizes/chemistry/2013/press-release/`.

[96] R. Verfürth. "A posteriori error estimates for nonlinear problems. Finite element discretizations of elliptic equations". In: *Mathematics of Computation* 62.206 (1994), pp. 445–475.

[97] R. Verfürth. "A posteriori error estimation and adaptive mesh-refinement techniques". In: *Journal of Computational and Applied Mathematics* 50.1-3 (1994), pp. 67–83.

[98] R. Verfürth. *A Posteriori Error Estimation Techniques for Finite Element Methods*. Oxford University Press, Apr. 2013.

[99] M. J. Wazwaz. "A posteriori error estimates for elliptic partial differential equations in the finite element method". In: (2017).

[100] A. M. Winslow. *Adaptive-mesh zoning by the equipotential method*. Tech. rep. Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States), 1981.

[101] F. Xu and Q. Huang. "Cascadic adaptive finite element method for nonlinear eigenvalue problem based on complementary approach". In: *Journal of Computational and Applied Mathematics* 372 (July 2020), p. 112720.

[102] D. M. Young. *Iterative solution of large linear systems*. Elsevier, 2014.

[103] H. Yserentant. "The Electronic Schrödinger Equation". In: *Regularity and Approximability of Electronic Wave Functions*. Springer, 2010, pp. 51–58.

[104] E. Zeidler. *Nonlinear functional analysis and its applications: II/B: nonlinear monotone operators*. Springer Science & Business Media, 2013.

[105] G. M. Zhislin. "A study of the spectrum of the Schrödinger operator for a system of several particles". In: *Trudy Moskovskogo Matematicheskogo Obshchestva* 9 (1960), pp. 81–120.

[106] G. M. Zhislin and A. G. Sigalov. "The spectrum of the energy operator for atoms with fixed nuclei on subspaces corresponding to irreducible representations of the group of permutations". In: *Izvestiya Rossiiskoi Akademii Nauk. Seriya Matematicheskaya* 29.4 (1965), pp. 835–860.

[107] A. Zhou. "Finite dimensional approximations for the electronic ground state solution of a molecular system". In: *Mathematical methods in the applied sciences* 30.4 (2007), pp. 429–447.

[108] T. G. Zieli. *Introduction to the finite element method*. Poland: Institute of Fundamental Technological Research of the Polish . . ., 1992.

[109] O. C. Zienkiewicz, R. L. Taylor, and J. Z. Zhu. *The finite element method: its basis and fundamentals*. Elsevier, 2005.

[110] "1 - Elementary Functions". In: *Table of Integrals, Series, and Products (Eighth Edition)*. Ed. by D. Zwillinger, V. Moll, I. Gradshteyn, and I. Ryzhik. Eighth Edition. Boston: Academic Press, 2014, pp. 25–62.

[111] T. Živković and H. Monkhorst. "Analytic connection between configuration–interaction and coupled-cluster solutions". In: *Journal of Mathematical Physics* 19.5 (2008), pp. 1007–1022.

Estimation d'erreur a posteriori pour des calculs de structure électronique par des méthodes ab initio et son application pour diminuer le cout de calcul
A posteriori error estimation for electronic structure calculations using ab initio methods and its application to reduce calculation costs

Abstract

The thesis is concerned with the error analysis of electronic structure calculation. The long term goal is to, in one hand, derive computable a posteriori error estimator for ab initio methods and, in the other hand, propose near-optimal computational cost strategy for the numerical calculation of those methods based on the a posteriori error estimation and the separation of the discretization and iteration error sources. In the first part of the thesis, we introduce a new well-posedness analysis for the single reference coupled cluster method based on the invertibility of the CC derivative. Under the minimal assumption that the sought-after eigenfunction is intermediately normalisable and the associated eigenvalue is isolated and non-degenerate, we prove that the continuous (infinite-dimensional) CC equations are always locally well-posed. Under the same minimal assumptions and provided that the discretization is fine enough, we prove that the discrete Full-CC equations are locally well-posed, and we derive residual-based error estimates with guaranteed positive constants. The second part of the thesis focus on the application of a posteriori error estimation to construct near-optimal path when approximating the solution of PDEs. We firstly apply a probabilistic method to explore an optimal path that minimizes the cost for the numerical resolution of linear and nonlinear elliptic source problems. Based on the analysis of those optimal paths, we propose two near-optimal strategies to achieve a given accuracy based on the error sources decomposition of the error estimator. Finally, we validate the feasibility of those near-optimal strategies by applying them to the numerical approximation of a nonlinear eigenvalue problem, i.e., the Gross-Pitaevskii equation.

Keywords: electronic structure theory, coupled cluster method, numerical analysis, non-linear functions, error estimate, Gross-Pitaevskii equation, residual decomposition.

Résumé

La thèse porte sur l'analyse des erreurs dans le calcul de la structure électronique. L'objectif à long terme est, d'une part, de dériver un estimateur d'erreur a posteriori calculable pour les méthodes ab initio et, d'autre part, de proposer une stratégie de coût de calcul quasi-optimale pour le calcul numérique de ces méthodes basée sur l'estimation d'erreur a posteriori et la séparation des sources d'erreur de discrétisation et d'itération. Dans la première partie de la thèse, nous introduisons une nouvelle analyse de bien posé pour la méthode de cluster couplé à référence unique basée sur l'inversibilité de la dérivée CC. Sous l'hypothèse minimale que la fonction propre recherchée est normalisable de façon intermédiaire et que la valeur propre associée est isolée et non dégénérée, nous prouvons que les équations CC continues (en dimension infinie) sont toujours bien posées localement. Sous les mêmes hypothèses minimales et à condition que la discrétisation soit suffisamment fine, nous prouvons que les équations CC discrètes sont localement bien posées, et nous dérivons des estimations d'erreur basées sur les résidus avec des constantes positives garanties. La deuxième partie de la thèse se concentre sur l'application de l'estimation d'erreur a posteriori pour construire un chemin quasi-optimal lors de l'approximation de la solution d'EDP. Nous appliquons d'abord une méthode probabiliste pour explorer un chemin optimal pour la résolution numérique de problèmes elliptiques linéaires et non linéaires en minimisant le coût de calcul. Sur la base de l'analyse de ces chemins optimaux, nous proposons deux stratégies quasi-optimales pour atteindre une précision donnée, basées sur la décomposition des sources d'erreur de l'estimateur d'erreur. Enfin, nous validons la faisabilité de ces stratégies quasi-optimales en les appliquant à l'approximation numérique du problème des valeurs propres, c'est-à-dire l'équation de Gross-Pitaevskii.

Mots clés : théorie de la structure électronique, méthode des clusters couplées, analyse numérique, fonction non linéaire, estimation d'erreur, équation de Gross-Pitaevskii, décomposition du residu.

Laboratoire Jacques-Louis Lions
Sorbonne Université – Campus Pierre et Marie Curie – 4 place Jussieu – 75005 Paris – France