Scalable Domain Decomposition Eigensolvers for Schrödinger Operators in Anisotropic Structures

—

Von der Fakultät für Mathematik, Informatik und Naturwissenschaften der RWTH Aachen University zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften genehmigte Dissertation

vorgelegt von

Lambert Theisen, M.Sc.
aus Bad Ems, Deutschland

Berichter: Prof. Dr. Benjamin Stamm
Prof. Dr. Arnold Reusken
Prof. Dr. Patrick Henning

Tag der mündlichen Prüfung: 13. Juni 2024

Diese Dissertation ist auf den Internetseiten der Universitätsbibliothek verfügbar.

Eigenständigkeitserklärung / Affidavit

Eidesstattliche Erklärung

Ich, Lambert Theisen, erklärt hiermit, dass diese Dissertation und die darin dargelegten Inhalte die eigenen sind und selbstständig, als Ergebnis der eigenen originären Forschung, generiert wurden. Hiermit erkläre ich an Eides statt

1. Diese Arbeit wurde vollständig oder größtenteils in der Phase als Doktorand dieser Fakultät und Universität angefertigt;
2. Sofern irgendein Bestandteil dieser Dissertation zuvor für einen akademischen Abschluss oder eine andere Qualifikation an dieser oder einer anderen Institution verwendet wurde, wurde dies klar angezeigt;
3. Wenn immer andere eigene- oder Veröffentlichungen Dritter herangezogen wurden, wurden diese klar benannt;
4. Wenn aus anderen eigenen- oder Veröffentlichungen Dritter zitiert wurde, wurde stets die Quelle hierfür angegeben. Diese Dissertation ist vollständig meine eigene Arbeit, mit der Ausnahme solcher Zitate;
5. Alle wesentlichen Quellen von Unterstützung wurden benannt;
6. Wenn immer ein Teil dieser Dissertation auf der Zusammenarbeit mit anderen basiert, wurde von mir klar gekennzeichnet, was von anderen und was von mir selbst erarbeitet wurde;
7. Teile dieser Arbeit wurden zuvor veröffentlicht und zwar in:
    - L. Theisen and B. Stamm. *A Scalable Two-Level Domain Decomposition Eigensolver for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains*. Submitted. 2023. DOI: 10.48550/arXiv.2311.08757. arXiv: 2311.08757 [cs, math].
    - B. Stamm and L. Theisen. "A Quasi-Optimal Factorization Preconditioner for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains". In: *SIAM J. Numer. Anal.* 60.5 (2022), pp. 2508–2537. DOI: 10.1137/21M1456005.

Aachen, Januar 2024

Lambert Theisen

Zusammenfassung

Diese Arbeit behandelt die Konstruktion und Analyse von skalierbaren Vorkonditionierungsstrategien für das lineare Schrödinger-Eigenwertproblem mit periodischen Potenzialen in anisotropen Strukturen. Da nur einige Dimensionen des Berechnungsgebiets gegen unendlich streben, wird die Eigenwertlücke zwischen dem ersten und zweiten Eigenwert verschwindend gering, was eine signifikante Herausforderung für iterative Löser darstellt.

Für diese iterativen Eigenwertlöser stellen wir daher eine quasi-optimale Strategie des Vorkonditionierens vor, die auf dem Prinzip der Spektralverschiebung-und-Invertierung beruht, sodass die iterativen Eigenwertlöser in einer konstanten Anzahl an Iterationen konvergieren. In der Analyse leiten wir eine analytische Faktorisierung der Eigenpaare her und nutzen die direktionale Homogenisierung, um das asymptotische Verhalten zu analysieren. Das resultierende, leicht zu berechnende, Einheitszellenproblem kann innerhalb des Spektralverschiebungs-Vorkonditionierers verwendet werden. Dieser Ansatz führt zu einer gleichmäßig beschränkten Anzahl an Eigenwertlöser-Iterationen. Numerische Beispiele veranschaulichen die Effektivität dieser quasi-optimalen Vorkonditionierungsstrategie, sofern direkte Löser verwendet werden, da die Verschiebestrategie, definitionsgemäß, zu einem kleineren Eigenwert für den resultierenden verschobenen Operator führt, was wiederum zu einer hohen Konditionszahl führt.

Weiterhin stellen wir einen zweistufigen Gebietszerlegungs-Vorkonditionierer für iterative lineare Löser vor, um genau dieses Problem zu lösen. Da die Berechnung der quasi-optimalen Verschiebung bereits die Lösung eines spektralen Zellenproblems als Grenz-Eigenfunktion bereitstellt, ist es naheliegend, diese als Generator zu verwenden, um einen Grobraum zu konstruieren. Tatsächlich ist es der Fall, dass der resultierende zweistufige additive Schwarz-Vorkonditionierer unabhängig von der Anisotropie des Gebiets ist, da wir eine Konditionszahl-Schranke unter Verwendung der Theorie der spektralen Grobräume erhalten, obwohl nur eine einzige Basisfunktion pro Teilgebiet benötigt wird. Wir stellen mehrere numerische Beispiele vor, die die Effektivität beider Methoden getrennt veranschaulichen, und kombinieren sie am Ende, um ihre kombinierte Skalierbarkeit zu zeigen.

Schlagwörter: periodische Schrödingergleichung, iterative Eigenwertlöser, Vorkonditionierer, asymptotische Eigenwertanalyse, Faktorisierungsprinzip, direktionale Homogenisierung, Gebietszerlegung, Grobräume, Finite-Elemente-Methode

MSC-Codes: 65N25, 65F15, 65N30, 65F10, 65N22, 65N55, 65F08, 35B27, 35B40

Abstract

This thesis presents the construction and analysis of scalable preconditioning strategies for the linear Schrödinger eigenvalue problem with periodic potentials in anisotropic structures. As only some dimensions of the computational domain expand to infinity, the resulting eigenvalue gap between the first and second eigenvalue vanishes, posing a significant challenge for iterative solvers.

For these iterative eigenvalue solvers, we provide a quasi-optimal shift-and-invert preconditioning strategy such that the iterative eigenvalue algorithms converge in constant iterations for different domain sizes. In its analysis, we derive an analytic factorization of the eigenpairs and use directional homogenization to analyze the asymptotic behavior. The resulting easy-to-calculated unit cell problem can be used within a shift-and-invert preconditioning strategy. This approach leads to a uniformly bounded number of eigensolver iterations. Numerical examples illustrate the effectiveness of this quasi-optimal preconditioning strategy if direct solvers are used since the shifting strategy, by definition, leads to a smaller eigenvalue for the resulting shifted operator, which, in turn, results in a high condition number.

We also provide a two-level domain decomposition preconditioner for iterative linear solvers to overcome this issue. As the calculation of the quasi-optimal shift already offered the solution to a spectral cell problem as limiting eigenfunction, it is logical to use it as a generator to construct a coarse space. Indeed, it is the case that the resulting two-level additive Schwarz preconditioner is independent of the domain's anisotropy since we obtain a condition number bound using the theory of spectral coarse spaces despite the need for only one basis function per subdomain for the coarse solver. We provide several numerical examples illustrating the effectiveness of both methods separately and combine them in the end to show their combined scalability.

Keywords: periodic Schrödinger equation, iterative eigenvalue solvers, preconditioner, asymptotic eigenvalue analysis, factorization principle, directional homogenization, domain decomposition, coarse spaces, finite element method

MSC Codes: 65N25, 65F15, 65N30, 65F10, 65N22, 65N55, 65F08, 35B27, 35B40

Acknowledgments

The research for this work was carried out between October 2019 and January 2024 at MathCCES/ACoM at the RWTH Aachen University and, most recently, at the NMH chair at the University of Stuttgart.

First of all, I would like to thank my supervisor, Beni. It was always impressive how quickly you got back into my topic and how we could continue the discussion immediately. The academic support was excellent, and you always cared a lot about my progress. I always particularly appreciated the freedom I had in my research. However, this freedom also brings a particular responsibility, for which I am grateful that you gave me a motivating boost when necessary to put me back on the right track. Looking back now, that was extremely helpful. I also learned a lot personally and how to approach life. Your support was also crucial during challenging times, including the COVID lockdown and our group's move from Aachen to Stuttgart. I never felt alone, and choosing Stuttgart and the Bahncard 100 lifestyle has been one of the most important decisions I've made in recent years, and I don't regret it. It has taken me a long way forward, personally. I also really enjoyed working with and taking responsibility for the organizational development of the new working group. All in all, I had a great time!

I would also like to thank Patrick Henning and Arnold Reusken for accepting the reviewer role and carefully reading this work, as well as Manuel Torrilhon and Christof Melcher for their willingness to complete the committee.

My sincere thanks also go to my colleagues from Aachen, like Hassan, Michael, Matthias, Jonas, Paul, Ullika, Aleksandr, Abhinav, Edilbert, Tamme, Vladimir, Daniel, Donat, Andrea and all the others I forgot. You were fantastic colleagues. Many thanks also to Manuel, who continued to offer me the opportunity to keep my premium office here in Aachen so that I could work from Aachen one day a week. I liked the ICE journeys to Stuttgart[1]. This fact was due in no small part to the people I met in Stuttgart, like Tiz, Michele, Louis, Gaspard, Paula, YingXing, Yao, Zahra, Thiago, Rafael, Markus, David, the Straußi people, and, of course, Brit! Thanks also to my personal friends and the students who have listened to me in exercises.

Special thanks to my parents, who always supported me and allowed me to do whatever I wanted. This absolute freedom, combined with the knowledge of unlimited support, is a privilege that I greatly appreciate.

The place of honor on this list goes to my girlfriend, Christin. Thank you for always being there for me, supporting me, and believing in me. I am very grateful to have you by my side and look forward to our future together.

[1] Although the ICE connection (`Aachen-Frankfurt`) $\wedge$ `(Frankfurt-Stuttgart)` is probably the most unreliable combination in Germany, based on personal sample size $n \approx 50$.

List of Publications and Scholarly Works

Since beginning my research in October 2019, I have been the corresponding author for three articles (two journal articles and one submitted preprint):

- [247]: L. Theisen and B. Stamm. *A Scalable Two-Level Domain Decomposition Eigensolver for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains*. Submitted. 2023. DOI: 10.48550/arXiv.2311.08757. arXiv: 2311.08757 [cs, math].
- [239]: B. Stamm and L. Theisen. "A Quasi-Optimal Factorization Preconditioner for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains". In: *SIAM J. Numer. Anal.* 60.5 (2022), pp. 2508–2537. DOI: 10.1137/21M1456005.
- [250]: L. Theisen and M. Torrilhon. "fenicsR13: A Tensorial Mixed Finite Element Solver for the Linear R13 Equations Using the FEniCS Computing Platform". In: *ACM Trans. Math. Softw.* 47.2 (2021), 17:1–17:29. DOI: 10.1145/3442378.

The following software packages (including milestone snapshots) have been made publicly available:

- [249]: L. Theisen and B. Stamm. *ddEigenLab.Jl: Domain-Decomposition Eigenvalue Problem Lab (v0.3)*. Zenodo. 2023. DOI: 10.5281/zenodo.10121779.
- [248]: L. Theisen and B. Stamm. *ddEigenLab.Jl: Domain-Decomposition Eigenvalue Problem Lab (v0.2)*. Zenodo. 2022. DOI: 10.5281/zenodo.6576197.
- [251]: L. Theisen and M. Torrilhon. *fenicsR13: A Tensorial Mixed Finite Element Solver for the Linear R13 Equations Using the FEniCS Computing Platform (v1.4)*. Zenodo. 2020. DOI: 10.5281/zenodo.4172951.

The following theses have been supervised[2]:

- [112]: L. Fiorenza. "Preconditioning in Steepest Descent Methods for Discretized Elliptic Eigenvalue Problems". Ongoing. Bachelor Thesis. University of Stuttgart, 2024.
- [43]: S. Berger. "Density Operator in Eigenvalue Problems with Application in Manifold Interpolation". Bachelor Thesis. RWTH Aachen University, 2022.

[2]The list does not contain multiple, but, for the topic of this work unrelated, seminar theses at the University of Stuttgart. Also, note that the work in [112] is ongoing at the time of writing.

- [202]: C. Müller, M. Geratz, C. Heger, and J. Meyer. “Evaluation and Implementation of Schrödinger-Type Eigenvalue Problems in Long Rectangular Domains Using the Finite Element Method”. CES Project Thesis. RWTH Aachen University, 2021.
- [46]: H. Borchardt. “Iterative Domain Decomposition Methods for Eigenvalue Problems”. Master Thesis. RWTH Aachen University, 2020.
- [176]: A. Kristof. “Using a Spectral Inference Network to Solve the Time-Independent Schrödinger Equation for a Two-Dimensional Hydrogen Atom”. CES Seminar Thesis. RWTH Aachen University, 2020.

A diverse collection of material and computational notebooks used for teaching purposes can also be found in the GitHub repository `@lamBOOO/teaching`[3].

[3] See `https://github.com/lamBOOO/teaching` or `https://thsn.dev`.

Authorship Declaration and Transparency Statement

As required by the doctoral regulations ("Promotionsordnung vom 27.09.2010 in der Fassung der siebten Ordnung zur Änderung der Promotionsordnung vom 07.09.2018, §5 (6)"), I note the following:
The content in Chapter 3 has been published in the article [239] and the content in Chapter 4 has been submitted and published as a preprint [247]:

- [239]: B. Stamm and L. Theisen. "A Quasi-Optimal Factorization Preconditioner for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains". In: *SIAM J. Numer. Anal.* 60.5 (2022), pp. 2508–2537. DOI: 10.1137/21M1456005
- [247]: L. Theisen and B. Stamm. *A Scalable Two-Level Domain Decomposition Eigensolver for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains*. Submitted. 2023. DOI: 10.48550/arXiv.2311.08757. arXiv: 2311.08757 [cs, math]

I declare that I majorly contributed to the articles and hold the corresponding authorship for both. Specifically, following the CRediT classification[4], my contributions included conceptualization, methodology, software, validation, formal analysis, investigation, resources, data curation, writing (original draft), writing (review & editing), and visualization. My co-author, Benjamin Stamm, contributed to the conceptualization, methodology, writing (review & editing), supervision, project administration, and funding acquisition. Both chapters represent the original articles listed above with minor contextual changes such as renaming variables, omitting the word "paper", and similar adjustments.
I also follow the DFG recommendation ("Stellungnahme des Präsidiums der Deutschen Forschungsgemeinschaft (DFG) zum Einfluss generativer Modelle für die Text- und Bilderstellung auf die Wissenschaften und das Förderhandeln der DFG, September 2023") and hereby declare the use of the following technologies, which contain or, as of the time of writing, could contain generative AI or related models: Google Search, Grammarly, DeepL Translator, ChatGPT, Google Translator, LanguageTool Translator, and GitHub Copilot. These tools have been used for tasks such as automatic formatting, copy editing, word translations, punctuation correction, spell checking, grammar checking, documentation generation of source code, source code auto completion, and general usage as search engines (if applicable).

Aachen, Januar 2024, Lambert Theisen

[4] See [49].

Contents

List of Figures

List of Tables

1 Introduction

Let us start with an observation from our daily lives. Assume that a person writes one page per day, then writing a short book with ten pages takes ten days. If ten persons want to write another book with a hundred pages, it should still take only ten days – in theory. In reality, however, it depends on the circumstances and the *scalability* of the task: Is the book a collection of ten separate chapters with ten pages each? Then, ten days seem reasonable. Is the book, however, meant to be the script for the next blockbuster movie? Then, it will take much longer since the separate writers depend on each other's work and must communicate.

Sometimes, making a clever decision before starting the task can make a massive difference. For example, painting a wall in white color together with a friend is scalable. However, we may want to paint the wall with two sections in two different colors, with 50 percent of the area in each color. Splitting the painting sections, i.e., the workload in that case, based on the desired color distribution, will speed up the task since everyone can work independently without switching the brushes. Thus, analyzing, if necessary, adapting problems is crucial to make them scalable.

1.1 Motivation and Context

In science and engineering, we often want to solve problems of size bigger than possible on a single computer or core. Thus, we must split the problem into smaller pieces and solve them independently. In the context of numerical simulation of a partial differential equation, this process is called *domain decomposition* (DD), where the domain $\Omega \subset \mathbb{R}^n$ is the region on which the equation, the model, is defined – just like the painting wall in the example above. The initial idea[1] for the DD method goes back to Schwarz [231] in 1870. The technique was initially formulated as a sequential algorithm with the paradigm *"solve left, then use the data to solve right"*. A parallel version with the adapted paradigm *"solve all domains simultaneously using the old data"* was proposed by Lions [192] when computers and large-scale computing clusters became more and more available. Having recognized the potential of such methods with the availability of hundreds, thousands, or even millions[2] of cores (nowadays), it

[1] Although, of course, with no initial intention for parallel computing.

[2] Compare with the 62nd, November 2023, edition of the TOP500 list (`www.top500.org`) that lists over eight million cores for the `Frontier` (rank 1 in the list) system.

was, however, noticed [152] that the *parallel efficiency* will decrease if the number of subdomains is increased and the scalability is lost.

In some cases, however, there are no issues when increasing the system size, and the number of iterations for the solution algorithm remains constant. For elliptic operators with Dirichlet boundaries, this phenomenon was observed for chain-like structures [65] in computational chemistry, analyzed in [86, 87, 88] with extensions in [90, 91, 136], and extends to the plane-like case [224]. In these cases, there is no need to use a coarse correction within a two-level framework. Analyzing this weak scalability for anisotropic structures, where each point within Ω has a fixed distance to the boundary when the domain size is expanded, is one of the main topics of this thesis. However, instead of PDE-based linear source problems, we want to consider the corresponding eigenvalue problem in these domains. Even if we choose the same kind of operators, i.e., second-order elliptic differential operators, the situation changes drastically for the eigenvalue problem. It requires different solution strategies to retain scalability.

Considering eigenvalue problems is not just based on mathematical curiosity. In computational chemistry, eigenvalue problems play a significant role in describing the electronic structure of molecules. We restrict ourselves to a related class of model problems, namely, the search for the first, i.e., lowest eigenpair, of linear periodic Schrödinger operators, and ask the question:

- Do we keep the one-level scalability of iterative eigenvalue solvers on expanding anisotropic structures, or does it change compared to the linear source problem?

We want to answer this question directly with a *no* since the convergence of the iterative eigenvalue algorithm fundamentally differs from iterative solvers for linear systems, to which the DD method belongs[3]. This difference is because it is no longer the condition number of the matrix but the ratio of the first to the second eigenvalue that determines the convergence rate. This ratio is related to the *spectral gap* – the difference between the smallest eigenvalues – and vanishes in the limit of anisotropically expanding domains. Thus, as a result, the convergence rate will become arbitrarily bad, and the number of iterations needed to solve the problem will tend to infinity.

As part of this thesis, we provide a strategy for solving this *collapsing gap problem* based on a quasi-optimal shifting approach. Although this shifting leads to new challenges, e.g., ill-conditioning of the resulting shifted operators, it yet allows us to make a fundamental observation that goes beyond the specific applications within this thesis: There is an essential connection between the shift-and-invert (SI) preconditioning of eigenvalue solvers and the treatment of the resulting ill-conditioned systems, e.g., by using spectral coarse in the context of DD methods. It is possible to incorporate the asymptotic information of the first eigenvalue and the limiting behavior of the first eigenfunction to construct a coarse space. This observation is the key to solving the resulting ill-conditioned linear systems. Further contributions will be outlined in the following.

[3] This behavior was observed in [46].

1.2 Goal and Objectives

Recalling the previous section, the general goal of this thesis is to *develop and analyze scalable iterative eigenvalue algorithms for the class of linear periodic Schrödinger operators on expanding anisotropic domains*. To achieve this goal, we will define the following objectives:

1. **Analysis of the behavior of the first eigenpair**: For periodic Schrödinger operators in expanding anisotropic structures, the first eigenvalue converges in the limit. Analyzing this behavior is the first step to resolving the collapsing gap problem.

2. **Development and analysis of a shift-and-invert preconditioner**: For iterative eigenvalue solvers, we will use the asymptotic first eigenvalue as a shift to overcome the convergence issue. The next objective will be to analyze the resulting shifted fundamental ratio and the convergence rate of SI-preconditioned iterative eigenvalue solvers.

3. **Construction and analysis of a scalable two-level DD method**: To achieve numerical scalability when dealing with the resulting ill-conditioned systems, the DD methods must use a coarse correction, or coarse space, in a two-level fashion. Thus, the final objective will be to construct a two-level DD method for the shifted linear systems and combine everything into one scalable iterative eigenvalue algorithm. The scalability will be quantified by analyzing the resulting condition number of the preconditioned matrix.

Furthermore, all the above-listed objectives will be complemented by numerical examples that illustrate the theoretical results and show the performance of the proposed methods for practical applications. Ensuring reproducibility by publicly providing all source codes can also be seen as an objective. Since the above-listed objectives require diverse concepts within different fields, part of this thesis will also review all necessary techniques and ideas to understand better how they are combined to achieve the objectives.

1.3 Outline and Contributions

Considering the hierarchical objectives listed in the previous Section 1.2, the overall thesis structure is linear, and the developed methods and, thus, chapters build on each other. Nevertheless, we recall the required concepts when necessary and motivate each chapter independently. The thesis is structured into three parts, which we briefly outline and synthesize.

In Chapter 2, we introduce and review all models, concepts, and computational tools needed for the following chapters. Starting with a general introduction to computational chemistry, we review the essential models used in electronic structure theory. This overview allows us to introduce the linear Schrödinger eigenvalue problem

as a simplified model problem while showing the connection to the original application. Considering the modeling-related big picture, we move on to its numerical treatment using iterative eigenvalue solvers. We review the basic concepts of such iterative methods, discuss different variants, and show their challenges for problems with a collapsing fundamental gap. The transition from iterative eigenvalue to iterative linear solvers is drawn by introducing the DD method. We review the basic concepts, historical remarks, and variants. We focus on the additive Schwarz methods for the latter since they are the most relevant for the following chapters. Finally, we introduce the concept of coarse spaces focusing on spectral constructions, which will play an essential role in the thesis.

In Chapter 3, we analyze the behavior of the first eigenpair of the linear Schrödinger equation in abstract d-dimensional anisotropic structures when the domain size L, for some $p = d - q$ dimensions, tends to infinity. In the process, an essential result is extended to the directional case, namely the factorization of the m-th eigenfunction $\phi_L^{(m)}$ into an easier-to-calculate unit cell eigenfunction φ and a remainder $u^{(m)}$, i.e.[4], $\phi_L^{(m)} = \varphi \cdot u^{(m)}$. Although, in our case, it is mainly used as an auxiliary tool in the analysis, it stands on its own. It allows us to characterize the spectral properties of the operator. The factorization is then used to transform the problem into an equivalent problem on the unit cell. This process creates a homogenization problem due to the rescaling in the domain length L. Once this is recognized, it remains to generalize the existing analysis to the directional case, which, after some technicalities, allows to characterize[5] the behavior of the eigenvalue as $\lambda_L^{(m)} = \sigma + (1/L^2)(\nu^{(m)} + \mathcal{O}(1/L))$ in which $\nu^{(m)}$ is the eigenvalue of a homogenized limit eigenvalue problem. Also, the corresponding weak limit for the eigenfunctions can be shown, achieving the first objective in Section 1.2. The general result can then be used to extract the asymptotic limit of the first eigenvalue. Using this limit as a shifting parameter for iterative eigenvalue solvers leads to a quasi-optimal method since the number of iterations is optimal, up to a multiplicative constant. Technically, this is shown by proving[6] that the fundamentally shifted ratio $r_L(\sigma) := |\lambda_L^{(1)} - \sigma|/|\lambda_L^{(2)} - \sigma|$ is uniformly bounded above by a constant $C < 1$ in the limit. However, although we have found a solution to the collapsing gap problem, thus having fulfilled the second objective in Section 1.2, we face new challenges.

These newly created ill-conditioning issues of the shifted linear systems are addressed in Chapter 4 to achieve the third objective in Section 1.2. We propose a two-level DD method that uses a coarse space to solve the resulting ill-conditioned systems. The proposed method becomes numerically scalable if the coarse space contains the periodic extensions of the unit cell functions φ (see above). This remarkable observation highlights the connection between the SI-preconditioning of eigenvalue solvers and preconditioning for linear solvers if the DD method is, e.g., used as a Krylov acceleration. Also, for the numerical treatment, this has some handy implications,

[4] See the Theorem 3.1 for a more detailed version that also includes further sub-factorizations.
[5] A precise statement is given in Theorem 3.2.
[6] See the Corollary 3.1.

namely that the coarse space components are already computed and that there is only one basis function per subdomain, which makes the treatment of the second level much more accessible. Technically, the periodic factorization strategy belongs to the category of spectral coarse spaces. Having recognized this relation, and under some additional symmetry assumptions, a stable decomposition property can be shown[7], which is one of the essential ingredients for the existing abstract theory and ultimately leads to a condition number bound[8], $\kappa(\boldsymbol{M}_{\mathrm{AS},2}^{-1}\boldsymbol{A}_{\sigma}) \leq C$ for some constant $C > 0$, independent of the domain size L. However, numerical results also suggest the same performance for the general case. We conclude the chapter with an observation, namely that the fusion of loops within an inner-outer eigensolver is beneficial to reducing the total number of inner iterations.

In the last Chapter 5, we summarize the results, highlight areas of improvement, and give perspectives on future work. Since the thesis is based on a mixture of different concepts and techniques that all contribute to the same goal, we discuss implications and future directions within the field by going a step back and discussing ideas for the abstract case of a general eigenvalue problem, depending on some arbitrary parameter ϵ, whose eigenvalue gap vanishes in the limit of $\epsilon \to 0$. Also, some ideas on connecting the field of eigenvalue preconditioning to the preconditioning of linear solvers are presented. The thesis then concludes with a list of references.

[7] See the Theorem 4.1.

[8] The condition number bound fits into the existing results form the spectral coarse space theory and is given in Theorem 4.2.

2 Preliminaries, the Big Picture, and an Overview of Methods

This chapter reviews the quantum many-body problem, the electronic structure calculation, and all related aspects to embed our results into the bigger picture. We then introduce the linear Schrödinger equation and the discretization methods we use to solve it. Finally, we overview iterative eigenvalue algorithms and domain decomposition methods.

2.1 Electronic Structure Theory

We start by overviewing the relevant parts of (non-relativistic) quantum mechanics (QM) required for our context in quantum and computational chemistry. For a more detailed introduction, we refer, e.g., to the classical books [84, 194, 199, 200, 203, 232, 242] and [56, 95, 133, 181, 195] for the mathematical perspective. We also keep the presentation compact by ignoring (for this work) unimportant aspects like spin or relativistic effects.

As a starting point, let us consider the time-dependent Schrödinger equation as

$$i\hbar\frac{\partial\Psi}{\partial t} = H\Psi, \tag{2.1}$$

in which $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$ is a vector of spatial coordinates, $t \in \mathbb{R}$ denotes the time, $\Psi(\boldsymbol{x}, t)$ is the wave function, i is the imaginary unit, $\hbar$ is the reduced Planck constant while the operator H is the so-called *Hamiltonian* operator. The Eq. (2.1) and an initial condition, $\Psi(\boldsymbol{x}, 0) = \Psi_0(\boldsymbol{x})$, form an *initial value problem* (IVP) for the wave function Ψ. The Hamiltonian H is a linear operator that describes the system's total energy. For example, the Hamiltonian for a single particle of mass m in a potential V in $d = n = 3$ dimensions is given [56] by

$$H_{\text{sp}} = -\frac{\hbar^2}{2m}\Delta_{\boldsymbol{x}} + V(\boldsymbol{x}, t), \tag{2.2}$$

in which $\Delta_{\boldsymbol{x}} := \sum_{i=1}^{3} \partial^2/\partial x_i^2$ is the Laplace operator for all spatial dimensions. Equations involving operators similar to Eq. (2.2) are all referred to as *linear Schrödinger equations* and serve as the main model problems throughout this work. Examples of the external potential V in Eq. (2.2), see [133], include, e.g., the free flight or free motion with $V(\boldsymbol{x}, t) = 0$, a wall with $V(\boldsymbol{x}, t) = 0$ on one side and $V(\boldsymbol{x}, t) = \infty$ on the other

side, implying $\Psi = 0$ behind the wall, the harmonic oscillator with $V(\boldsymbol{x}, t) = \frac{1}{2}m\omega^2\boldsymbol{x}^2$ with the angular frequency ω, the hydrogen atom with $V(\boldsymbol{x}, t) = -\frac{e^2}{4\pi\epsilon_0}\frac{1}{\|\boldsymbol{x}\|}$ with the dielectric permittivity ϵ_0 and the elementary charge e, a particle in a box Ω with $V(\boldsymbol{x}, t) = 0$ if $\boldsymbol{x} \in \Omega$ and $V(\boldsymbol{x}, t) = \infty$ else, and a periodic potential with $V(\boldsymbol{x}, t) = V(\boldsymbol{x} + \boldsymbol{i}, t)$ for all, w.l.o.g., $\boldsymbol{i} \in \mathbb{Z}^n$. The case of periodic potentials is of particular interest to this work, see Chapters 3 and 4, since it can model periodic lattices and, mathematically, has a well-behaved limit when the size of the domain Ω grows to infinity.

In general, the Eq. (2.1) is challenging to solve since H can be very complex, and the dimension of $\boldsymbol{x}$ is not limited to three. However, linearity yields the following.

- **Superposition principle**: If Ψ_1 and Ψ_2 are solutions of Eq. (2.1), so is the linear combination $\Psi = \alpha\Psi_1 + \beta\Psi_2$ for any $\alpha, \beta \in \mathbb{C}$ a solution to Eq. (2.1).

For a time-independent Hamiltonian, i.e., $H = H(\boldsymbol{x})$, knowing the eigenfunctions of H allows us to calculate the time dynamics of the system [56, 132]. Assuming Ψ_0 to be a normalized eigenfunction of H, i.e.,

$$H(\boldsymbol{x})\Psi_0(\boldsymbol{x}) = E\Psi_0(\boldsymbol{x}), \quad \|\Psi_0\| = 1, \tag{2.3}$$

with the eigenvalue (or energy) $E \in \mathbb{R}$, then the time evolution of Ψ is given [203, p23] by

$$\Psi(\boldsymbol{x}, t) = \Psi_0(\boldsymbol{x})e^{-iEt/\hbar}. \tag{2.4}$$

If, on the other hand, $\Psi_0(\boldsymbol{x})$ is a linear combination of multiple eigenfunctions $\{\psi_n(\boldsymbol{x})\}_{n\in\mathcal{N}}$ with corresponding eigenvalues E_n, i.e.,

$$\Psi_0(\boldsymbol{x}) = \sum_{n\in\mathcal{N}} \alpha_n\psi_n(\boldsymbol{x}), \tag{2.5}$$

then, using the superposition principle, the time evolution of Ψ is given by

$$\Psi(\boldsymbol{x}, t) = \sum_{n\in\mathcal{N}} \alpha_n\psi_n(\boldsymbol{x})e^{-iE_nt/\hbar}. \tag{2.6}$$

In general, solutions of the form

$$\Psi(\boldsymbol{x}, t) = \psi(\boldsymbol{x})e^{-iEt/\hbar}, \tag{2.7}$$

are called *stationary states*[1] and lead to a time-independent probability density $|\Psi(\boldsymbol{x}, t)|^2 = |\psi(\boldsymbol{x})|^2$ [92]. Plugging Eq. (2.7) into Eq. (2.1) and using a separation into an $\boldsymbol{x}$- and t-dependent part yields the *time-independent Schrödinger equation* [56, 200] as

$$H\psi_n = E_n\psi_n, \quad \|\psi_n\| = 1, \tag{2.8}$$

which is fundamental to understanding the properties of the considered system. Thus, finding the eigenpairs of H is a crucial problem in quantum mechanics and the description of molecular systems, as shown in the following.

[1] Sometimes, the term *standing waves* is also used in this context [68, 194], which allows us to imagine the analogy to two vibrating strings with a phase shift of $\pi/2$, representing the real and the imaginary part of Ψ, when applying, e.g., the "particle in a box" potential.

2.1.1 The Many-Body Description of Molecular Systems

Moving on to the many-body description in electronic structure calculations, we now consider a system of N interacting electrons with positions[2] $\boldsymbol{x} := (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$, $\boldsymbol{x}_i \in \mathbb{R}^3$, mass m_e and charges $-e$ together with M nuclei with positions $\bar{\boldsymbol{x}} := (\bar{\boldsymbol{x}}_1, \ldots, \bar{\boldsymbol{x}}_M)$, $\bar{\boldsymbol{x}}_i \in \mathbb{R}^3$, masses $m_{\mathrm{n},1}, \ldots, m_{\mathrm{n},M}$ and electric charges $z_1 e, \ldots, z_M e > 0$. The non-relativistic many-body (molecular) Hamiltonian for this system is again the sum of the kinetic energies of the electrons and nuclei. At the same time, the potential is the sum of Coulomb interactions between each pair of particles [195]. Thus, we have

$$H = T + V, \quad \text{where } T = T_\mathrm{e} + T_\mathrm{n} \text{ and } V = V_\mathrm{ee} + V_\mathrm{ne} + V_\mathrm{nn}. \tag{2.9}$$

The kinetic energy terms are now given by the sum of all particles with their respective coordinates, i.e.,

$$T_\mathrm{e} = -\sum_{i=1}^{N} \frac{\hbar^2}{2m_\mathrm{e}} \Delta_{\boldsymbol{x}_i}, \quad T_\mathrm{n} = -\sum_{i=1}^{M} \frac{\hbar^2}{2m_{\mathrm{n},i}} \Delta_{\bar{\boldsymbol{x}}_i}, \tag{2.10}$$

while the potential energy terms are provided by the sum over all pairs of particles (electron-electron, electron-nucleus, nucleus-nucleus), i.e.,

$$V_\mathrm{ee} = \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{e^2}{4\pi\epsilon_0} \frac{1}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2}, \quad V_\mathrm{ne} = -\sum_{i=1}^{N} \sum_{j=1}^{M} \frac{z_j e^2}{4\pi\epsilon_0} \frac{1}{\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j\|_2}, \tag{2.11}$$

$$\text{and } V_\mathrm{nn} = \sum_{i=1}^{M} \sum_{j=i+1}^{M} \frac{z_i z_j e^2}{4\pi\epsilon_0} \frac{1}{\|\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_j\|_2}. \tag{2.12}$$

No matter which system we look at (e.g., the water molecule (H_2O) with $M = 3$ nuclei, the caffeine molecule ($C_8H_{10}N_4O_2$) with $M = 24$, ...), we can always use the Hamiltonian (2.9) without the need for system-specific, empirical parameters or heuristics. Therefore, the model is based on *first principles* [56][3].

To get more convenient numerical values[4], it is widespread to consider the equation (2.9) in *atomic units* [195], meaning that

$$\hbar = 1, \quad m_\mathrm{e} = 1, \quad e = 1, \quad \epsilon_0 = 1/(4\pi). \tag{2.13}$$

Technically, this results from rescaling the length unit, i.e., $\boldsymbol{x} \mapsto a_0 \tilde{\boldsymbol{x}}, \bar{\boldsymbol{x}} \mapsto a_0 \tilde{\bar{\boldsymbol{x}}}$ with the Bohr radius $a_0 = (4\pi\epsilon_0\hbar^2)/(m_\mathrm{e} e^2)$ (see [84]). The resulting rescaled Hamiltonian in atomic units, after dropping the $\tilde{\cdot}$-notation immediately for the rescaled quantities,

[2] An alternative notation uses $\boldsymbol{r} = (\boldsymbol{r}_1, \ldots, \boldsymbol{r}_N)$ and $\boldsymbol{R} = (\boldsymbol{R}_1, \ldots, \boldsymbol{R}_M)$ for $\boldsymbol{x}$ and $\bar{x}$, see [56].

[3] The term *ab initio* (from the Latin *from the beginning*) is also used to describe the modeling using first principles, c.f., [95].

[4] According to [84], another reason is the more straightforward comparison of different calculations since in atomic units, they do not depend anymore on the current best values for the constants in Eq. (2.9).

is then given by

$$\begin{aligned} H = &- \sum_{i=1}^{N} \frac{1}{2} \Delta_{\boldsymbol{x}_i} - \sum_{i=1}^{M} \frac{1}{2M_{\mathrm{n,i}}} \Delta_{\bar{\boldsymbol{x}}_i} - \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{z_j}{\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j\|_2} \\ &+ \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{1}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2} + \sum_{i=1}^{M} \sum_{j=i+1}^{M} \frac{z_i z_j}{\|\bar{\boldsymbol{x}}_i - \bar{\boldsymbol{x}}_j\|_2}, \end{aligned} \tag{2.14}$$

where $M_{\mathrm{n},i}$ denotes the ratio of the i-th nuclei mass to the electron mass, i.e., $M_{\mathrm{n},i} = m_{\mathrm{n},i}/m_{\mathrm{e}}$. For the Hamiltonian (2.14), we observe the following difficulties:

- The Hamiltonian acts on a function of very high dimensionality $n = 3N + 3M$, which is extremely hard to solve, even for small molecules. For the water molecule (H_2O), e.g., we would have $N = 10$ and $M = 3$ leading to $n = 39$. Naively storing the wave function Ψ in a regular grid with 100 points per dimension would require $100^{39} = 10^{78}$ points – a number having almost as many digits than the number of estimated particles in the universe [260].
- The factors $1/M_{\mathrm{n},i}$ of the kinetic energy terms of the nuclei are tiny since the ratios $M_{\mathrm{n},i}$ of nuclei masses to electron masses is enormous. This mass difference leads to very different scales between the electrons and nuclei.

These observations motivate the so-called *Born–Oppenheimer* approximation [47], which – for our case, simplified – treats the nuclei as point-like, classical particles such that their positions $\bar{\boldsymbol{x}}$ are fixed parameters for the movement of the electrons[5]. This separation implies that the nuclei are classically modeled by Newton's second law of motion, assuming that the electrons adapt instantaneously to any new nuclei arrangement. This assumption can be seen as a model reduction step since the dimension is reduced from $3(N + M)$ to $3N$. Thus, in the Eq. (2.14), the kinetic energy terms in T_{n} are neglected, and the potential energy terms in V_{nn} are treated as constants – only affecting the eigenvalues of the Hamiltonian but not the eigenfunctions. These considerations lead to the *electronic Hamiltonian* [242], given by

$$H_{\mathrm{e}} = H - T_{\mathrm{n}} - V_{\mathrm{nn}} = - \sum_{i=1}^{N} \frac{1}{2} \Delta_{\boldsymbol{x}_i} - \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{z_j}{\|\boldsymbol{x}_i - \bar{\boldsymbol{x}}_j\|_2} + \sum_{i=1}^{N} \sum_{j=i+1}^{N} \frac{1}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2}. \tag{2.15}$$

With these approximations, only the electrons are considered quantum particles that are described by the electronic wave function [105] $\psi_{\mathrm{e}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N)$ with

$$\psi_{\mathrm{e}} \in \bigotimes_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C}), \tag{2.16}$$

[5] Actually, the nuclei and electron coordinates are separated and the wave function factorized, leading to a nested optimization problem involving a geometry optimization of the nuclei coordinates $\bar{\boldsymbol{x}}$, such that they lead to the lowest total energy when also considering the electronic solution with $\bar{\boldsymbol{x}}$ as parameters, see, e.g.,[59, Eq. (6.7)] or [181] and further discussions in, e.g., [56, 195].

where $\otimes$ denotes the tensor product, and $L^2(\mathbb{R}^3, \mathbb{C})$ is the Lebesgue space of square-integrable functions on $\mathbb{R}^3$ with values in $\mathbb{C}$. Sometimes, the notation $\psi_\mathrm{e} \in L^2(\mathbb{R}^{3N}, \mathbb{C})$ is also used [56] since functions in $\in L^2(\mathbb{R}^{3N}, \mathbb{C})$ can be approximated by functions in $\bigotimes_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C})$ [105]. The space is a Hilbert space with the inner product

$$\langle \psi, \phi \rangle := \int_{\mathbb{R}^3} \psi(\boldsymbol{x}_1 \dots \boldsymbol{x}_N)^* \phi(\boldsymbol{x}_1 \dots \boldsymbol{x}_N) \, \mathrm{d}\boldsymbol{x}_1 \cdots \mathrm{d}\boldsymbol{x}_N \quad \forall \psi, \phi \in \bigotimes_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C}), \tag{2.17}$$

where ψ^* denotes the complex conjugate of ψ. The norm of ψ is then given by $\|\psi\| = \langle \psi, \psi \rangle^{1/2}$. From a physical standpoint [59], the squared modulus $|\psi_\mathrm{e}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)|^2 = \psi_\mathrm{e}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)^* \psi_\mathrm{e}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$ of the wave function ψ_e is the probability density of finding the N indistinguishable electrons at positions $(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$. Thus, the electronic wave function ψ_e is normalized since the total probability of finding the electrons is one, i.e.,

$$\|\psi_\mathrm{e}\|^2 = \int_{\mathbb{R}^{3N}} |\psi_\mathrm{e}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)|^2 \, \mathrm{d}\boldsymbol{x}_1 \cdots \mathrm{d}\boldsymbol{x}_N = 1. \tag{2.18}$$

Furthermore, since electrons are *fermions*, the wave function ψ_e is anti-symmetric concerning the exchange of any two electrons, i.e., meaning that

$$\psi_\mathrm{e}(\dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_j, \dots) = -\psi_\mathrm{e}(\dots, \boldsymbol{x}_j, \dots, \boldsymbol{x}_i, \dots), \quad \forall i, j \in \{1, \dots, N\}. \tag{2.19}$$

The property (2.19) is a mathematical consequence[6] of the indistinguishability [59] of identical particles, i.e., no change in $|\psi_\mathrm{e}|^2$ when switching positions of two particles, and allows for deducing the *Pauli exclusion principle*

$$\psi_\mathrm{e}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) = 0 \quad \text{if } \boldsymbol{x}_i = \boldsymbol{x}_j \text{ for some } i \neq j. \tag{2.20}$$

Thus, the electronic state space $\mathcal{H}_\mathrm{e}$ is given by

$$\mathcal{H}_\mathrm{e} = \bigwedge_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C}), \tag{2.21}$$

meaning that we only keep anti-symmetrized tensor products [59] from $\bigotimes_{i=1}^{N} L^2(\mathbb{R}^3, \mathbb{C})$.

Remark 2.1 (The spin variable). In the complete picture, electrons have another variable, $\sigma_i \in \{-\frac{1}{2}, \frac{1}{2}\}$. This variable is called *spin* and would typically change the electronic wave function to become $\psi_\mathrm{e}(\boldsymbol{x}_1, \sigma_1, \dots, \boldsymbol{x}_N, \sigma_N)$ [59] leading, e.g., to a modified space $\mathcal{H}_\mathrm{e} = \bigwedge_{i=1}^{N} L^2(\mathbb{R}^3 \times \{-\frac{1}{2}, \frac{1}{2}\}, \mathbb{C})$ in Eq. (2.21). Although spin has important practical implications [181], we omit it in this work for simplicity and clarity since it is irrelevant to our presentation.

[6] Symmetric functions would be the other possible case, which defines *bosons* [56, 59].

2.1.2 The Electronic Ground State Problem

In electronic structure calculations, the main focus is typically the computation of the smallest eigenpair of H_e. The underlying physical interpretation uses the energy, $E(\psi) = \langle \psi, H_\mathrm{e}\psi \rangle$, and that the lowest energy state is the most stable state of the system. The problem of calculating this electronic *ground state* is thus the energy minimization problem [56, 59] given by

$$E_0 = \inf\{\langle \psi_\mathrm{e}, H_\mathrm{e}\psi_\mathrm{e} \rangle,\ \psi_\mathrm{e} \in \mathcal{H}_\mathrm{e},\ \|\psi_\mathrm{e}\| = 1\}. \tag{2.22}$$

Note that the electronic Hamiltonian (2.15) is a real-valued operator and thus does not mix real and imaginary parts [105]. Therefore, the real and the imaginary parts of the ground state wave function are themselves minimizers of Eq. (2.15) (up to normalization and provided to be non-trivial). Thus, it is sufficient to consider the real-valued case [59]. Also, to only consider states of finite energy [59], we additionally can restrict the space $\mathcal{H}_\mathrm{e}$ only to allow functions with L^2-integrable first derivative to give a proper meaning to the kinetic terms, i.e., replacing $\mathcal{H}_\mathrm{e}$ by $\bigwedge_{i=1}^{N} H^1(\mathbb{R}^3)$, which we assume in the following, on which H_e is self-adjoint [59], and where the Sobolev space $H^1(\mathbb{R}^3)$ contains real-valued functions with L^2-integrable first derivative.

Defining the Lagrangian $\mathcal{L} : \mathcal{H}_e \times \mathbb{R} \to \mathbb{R}$ of the variational problem (2.22) as $\mathcal{L}(\psi_\mathrm{e}, \lambda) := \langle \psi_\mathrm{e}, H_\mathrm{e}\psi_\mathrm{e} \rangle + \lambda(1 - \|\psi_\mathrm{e}\|^2)$ and calculating critical points by setting the first variation of $\mathcal{L}$ w.r.t. ψ_e and the derivative of $\mathcal{L}$ w.r.t. λ to zero, i.e., the first-order optimality conditions, we obtain the variational *Euler–Lagrange equation* of Eq. (2.22) [238]: Find $(\psi_\mathrm{e}, \lambda) \in \mathcal{H}_\mathrm{e} \times \mathbb{R}$, such that

$$\forall \phi \in \mathcal{H}_\mathrm{e} : \quad \langle H_\mathrm{e}\psi_\mathrm{e}, \phi \rangle = \lambda \langle \psi_\mathrm{e}, \phi \rangle. \tag{2.23}$$

The Eq. (2.23) is the *weak form* of the time-independent Schrödinger eigenvalue problem for the electronic wave function, while the Lagrange multiplier λ is the energy. The corresponding *strong form* is – again – the time-independent Schrödinger equation Eq. (2.8), now for the electronic Hamiltonian H_e, given by

$$H_\mathrm{e}\psi_\mathrm{e} = E_\mathrm{e}\psi_\mathrm{e}, \quad \|\psi_\mathrm{e}\| = 1. \tag{2.24}$$

Solving the weak form (2.23) or directly minimizing the Eq. (2.22) is still challenging due to the problem's high dimensionality and the combinatoric structure. Therefore, we need to introduce further approximations and solution methods, which we discuss in the following.

2.1.3 Approximation Methods for the Ground State Problem

Like in every other discipline, finding accurate but cheaper-to-compute models, simplifications, or approximations is a crucial aspect of electronic structure calculations. Although we can not give a complete overview of all approximation methods, we want to briefly overview the most common approaches to show the relation to the methods we present in this work. For a more detailed overview, we refer to the techniques listed in, e.g., [56, 59, 142, 181] and references therein.

The following families of approximations can characterize the most common [181] approximation methods:

- **Wave function methods** directly operate on the wave function ψ_e and the variational problem (2.22) by replacing the optimization space $\mathcal{H}_\mathrm{e}$ with a space with reduced dimension. However, the operator and, thus, the energy are kept in their original form. One example is the *Hartree–Fock* method (HF) that restricts $\mathcal{H}_\mathrm{e}$ to so-called *Slater determinants* to transform the N-body problem to N one-body problems coupled nonlinearly. Other approaches can then be used to improve the HF solution, e.g., the *configuration interaction* method (CI), the *coupled cluster* method (CC), or approaches based on perturbation theory.

- **Density functional methods** that reformulate the problem (2.22) in terms of the *electronic density* $\rho_{\psi_\mathrm{e}}(\boldsymbol{x}) = N \int_{\mathbb{R}^{3(N-1)}} |\psi_\mathrm{e}(\boldsymbol{x}, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)|^2 \,\mathrm{d}\boldsymbol{x}_2 \cdots \mathrm{d}\boldsymbol{x}_N$, which is a function in $\mathbb{R}^3$ and thus has a much lower dimension than ψ_e. For a region $\omega \subset \mathbb{R}^3$, $\int_\omega \rho_{\psi_\mathrm{e}}(\boldsymbol{x})/N \,\mathrm{d}\boldsymbol{x}$ is the probability finding *any* electron in ω. Using the electronic density is very intuitive since, in the end, we are interested in the distribution of all electrons rather than specific ones. This approach is also mathematically sound since the *density functional theory* (DFT) states that the ground state energy can be obtained by minimizing some density-functional. The *Kohn–Sham* method (KS) is the most common method in this family, which we will briefly discuss in the following sections. However, the modification of the problem comes with a price, namely the nonlinearization of the problem [163] and, even more challenging, the search for the unknown functional since, although the theory states the existence of such a functional, its exact form is unknown [104].

In [59], a fascinating alternative perspective for the above list is presented: Methods of the first family tackle the variational problem by finite-dimensional approximations of the infinite-dimensional set of test functions (such as the finite element method, see Section 2.1.4.2); methods of the second kind, on the other hand, approximate the operator or the unknown energy functional itself (like the finite difference method). When reconsidering the Eq. (2.22), the only remaining aspect of further approximation (besides function space and operator) could be the evaluation of the energy integrals $\langle \psi_\mathrm{e}, H_\mathrm{e}\psi_\mathrm{e}\rangle$, which, if done stochastically, leads to variational Monte Carlo methods [59].

2.1.3.1 The Hartree–Fock Method

Following Hartree's motivation in [135, Sec. 1.8], if we assume no interaction between the electrons, i.e., $V_\mathrm{ee} = 0$ in Eq. (2.15), the electronic Hamiltonian H_e is separable. Ignoring, for now, the any-symmetry requirement, the separation of variables for the ground state problem (2.24) would lead to a solution that is a product wave function[7]

[7] Sometimes also called *Hartree product* [195].

of single-particle functions

$$\psi(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) = \prod_{i=1}^{N} \psi_i(\boldsymbol{x}_i) = \psi_1(\boldsymbol{x}_1) \cdots \psi_N(\boldsymbol{x}_N). \tag{2.25}$$

Another perspective is to use the product wave function form as an Ansatz for the space[8] $\bigotimes_{i=1}^{N} H^1(\mathbb{R}^3)$, resulting in the finite-dimensional subspace consisting of functions in the product wave function form (2.25). However, since we consider electronic wave functions, the anti-symmetry holds for electrons. Respecting this constraint can be done by using the *Slater determinants* [233] of the single-particle functions ϕ_i, i.e., following the notation[9] from [59, 181],

$$\psi_{\mathrm{e}}(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N) = \frac{1}{\sqrt{N!}} \det\left(\phi_i(\boldsymbol{x}_j)\right) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \phi_1(\boldsymbol{x}_1) & \phi_1(\boldsymbol{x}_2) & \cdots & \phi_1(\boldsymbol{x}_N) \\ \phi_2(\boldsymbol{x}_1) & \phi_2(\boldsymbol{x}_2) & \cdots & \phi_2(\boldsymbol{x}_N) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_N(\boldsymbol{x}_1) & \phi_N(\boldsymbol{x}_2) & \cdots & \phi_N(\boldsymbol{x}_N) \end{vmatrix}, \tag{2.26}$$

where the functions $\phi_i \in H^1(\mathbb{R}^3)$ are called *molecular orbitals* [181], are chosen to be $L^2(\mathbb{R}^3)$-orthonormal such that the factor $1/\sqrt{N!}$ in Eq. (2.26) ensures normalization of ψ_{e}. We define the set[10] containing all functions of the form (2.26) as

$$V^{\mathrm{HF}} = \left\{ \psi_{\mathrm{e}} = \frac{1}{\sqrt{N!}} \det\left(\phi_i(\boldsymbol{x}_j)\right) \;\middle|\; \phi_i \in H^1(\mathbb{R}^3),\ \langle \phi_i, \phi_j \rangle = \delta_{ij},\ i,j \in \{1, \dots, N\} \right\}. \tag{2.27}$$

Rewriting [199] the expression $\langle \psi_{\mathrm{e}}, H_{\mathrm{e}} \psi_{\mathrm{e}} \rangle$ while using that ψ_{e} is in V_{HF} leads to the Hartree–Fock energy minimization problem [56, 95] as

$$\begin{aligned} E_0^{\mathrm{HF}} &= \inf\{\langle \psi_{\mathrm{e}}, H_{\mathrm{e}} \psi_{\mathrm{e}} \rangle,\ \psi_{\mathrm{e}} \in V^{\mathrm{HF}}\} \\ &= \inf\{E^{\mathrm{HF}}(\boldsymbol{\phi}),\ \boldsymbol{\phi} = (\phi_1, \dots, \phi_N) \in (H^1(\mathbb{R}))^N, \langle \phi_i, \phi_j \rangle = \delta_{ij}\}, \end{aligned} \tag{2.28}$$

where the Hartree–Fock energy functional $E^{\mathrm{HF}}(\boldsymbol{\phi})$ is given [59] by

$$\begin{aligned} E^{\mathrm{HF}}(\boldsymbol{\phi}) =& \frac{1}{2} \sum_{i=1}^{N} \int_{\mathbb{R}^3} |\nabla \phi_i(\boldsymbol{x})|^2 \, \mathrm{d}\boldsymbol{x} + \int_{\mathbb{R}^3} \rho_{\boldsymbol{\phi}}(\boldsymbol{x}) V_{\mathrm{nuc}}(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \\ &+ \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\rho_{\boldsymbol{\phi}}(\boldsymbol{x}) \rho_{\boldsymbol{\phi}}(\boldsymbol{x}')}{\|\boldsymbol{x} - \boldsymbol{x}'\|_2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{x}' - \frac{1}{2} \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{|\gamma_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{x}')|^2}{\|\boldsymbol{x} - \boldsymbol{x}'\|_2} \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{x}', \end{aligned} \tag{2.29}$$

with the electronic density $\rho_{\boldsymbol{\phi}}(\boldsymbol{x}) := \sum_{i=1}^{N} |\phi_i(\boldsymbol{x})|^2 = \rho_{\psi_{\mathrm{e}}}(\boldsymbol{x})$ for $\psi_{\mathrm{e}} \in V^{\mathrm{HF}}$, the *(one-electron) density matrix* $\gamma_{\boldsymbol{\phi}}(\boldsymbol{x}, \boldsymbol{x}') := \sum_{i=1}^{N} \phi_i(\boldsymbol{x}) \phi_i(\boldsymbol{x}')$ and the nuclear Coulomb potential $V_{\mathrm{nuc}}(\boldsymbol{x}) := -\sum_{i=1}^{M} z_i / \|\boldsymbol{x} - \bar{\boldsymbol{x}}\|_2$ [59].

[8] Recall that in the most general case, an arbitrary element of the tensor product space is a converging infinite series of such Hartree products, as discussed in [181]. We also refer to, e.g., [137] for more details on the subspace of anti-symmetrized products.

[9] Sometimes, the notation $\psi_{\mathrm{e}} = |\phi_1, \dots, \phi_N\rangle$ is also used to indicate a Slater determinant [242].

[10] The set of Slater determinant wave functions V^{HF} is not a vector space, as, e.g., the sum of two Slater determinants in not necessarily a Slater determinant.

Since $V^{\mathrm{HF}} \subset \mathcal{H}_{\mathrm{e}}$, we have that the Hartree–Fock ground state energy is an upper bound for the exact ground state energy, i.e., $E_0 \leq E_0^{\mathrm{HF}}$. Under the assumption $N \leq Z := \sum_{i=1}^{M} z_i$ (meaning a neutral or positively charged system), there exists a ground state $\boldsymbol{\phi}^0 = (\phi_1^0, \ldots, \phi_N^0) \in (H^1(\mathbb{R}))^N$, see the works by Lieb [188] and Lions [189], that satisfy the *Hartree–Fock equations* [181]

$$\begin{cases} F_{\boldsymbol{\phi}}\phi_i(\boldsymbol{x}) = \varepsilon_i \phi_i(\boldsymbol{x}) \\ \langle \phi_i(\boldsymbol{x}), \phi_j(\boldsymbol{x}) \rangle = \delta_{ij} \end{cases}, \tag{2.30}$$

which are the Euler–Lagrange equations to Eq. (2.28), followed by a unitary transformation of $\boldsymbol{\phi}$ to diagonalize the Lagrange-multipliers [181]. The *Fock operator* in Eq. (2.30), following the notation[11] of [104, 105], is given by

$$\begin{aligned} \varphi(\boldsymbol{x}) \mapsto F_{\boldsymbol{\phi}}\varphi(\boldsymbol{x}) := &-\frac{1}{2}\Delta_{\boldsymbol{x}}\varphi(\boldsymbol{x}) + V_{\mathrm{nuc}}(\boldsymbol{x})\varphi(\boldsymbol{x}) + \sum_{i=1}^{N} \int_{\mathbb{R}^3} \frac{|\phi_i(\boldsymbol{x}')|^2}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2} \,\mathrm{d}\boldsymbol{x}' \varphi(\boldsymbol{x}) \\ &- \sum_{i=1}^{N} \int_{\mathbb{R}^3} \frac{\phi_i(\boldsymbol{x}')\varphi(\boldsymbol{x}')}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2} \,\mathrm{d}\boldsymbol{x}' \phi_i(\boldsymbol{x}). \end{aligned} \tag{2.31}$$

In Eq. (2.31), there is still some correlation between the electrons present, in contrast to the initial discussion of the chapter in Eq. (2.25). The underlying reason is the anti-symmetry requirement [59] we imposed on the wave function ψ_{e}. The HF equations (2.30) are challenging since they are nonlinear eigenvalue problems and contain non-local operators. However, as discussed in Section 2.1.4, the HF method is still very successful in practice since the dimension is reduced to N three-dimensional problems, allowing for more standard discretization. The nonlinear nature also required iterative (*self-consistent*) solution strategies, which we will discuss in Section 2.1.5.

The HF equations can also be used as a starting point for more accurate calculations, and such methods are called *post–Hartree–Fock methods* (see [59] for an overview). In the Configuration interaction method [59], e.g., when discretizing the operator $F_{\boldsymbol{\phi}}$ with $N_{\mathrm{b}} > N$ basis functions (see Section 2.1.4) and fully-diagonalizing the resulting matrix, we also obtain $N_{\mathrm{b}} - N$ additional, *virtual orbitals*, that can be used to build up other Slater determinants. The wave function is then written as a linear combination, with $\mathcal{I} \subset \{1, \ldots, N_{\mathrm{b}}\}, |\mathcal{I}| = N$, to construct

$$\psi_{\mathrm{e}}(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \sum_{\text{all or some } \mathcal{I}} c_{\mathcal{I}} \det\left(\phi_i(\boldsymbol{x}_j)_{i\in\mathcal{I}, j\in\{1,\ldots,N\}}\right). \tag{2.32}$$

The ansatz (2.32) is then again used within a variational principle to minimize the energy of H_{e}. Due to its combinatorial nature, this approach quickly becomes very expensive, so the additional orbitals to include and their combination are chosen carefully.

[11] An alternative notation, see, e.g., [59, 181], writes the operator $F_{\boldsymbol{\phi}}$ using the convolution, defined by $(f \star g)(\boldsymbol{x}) := \int_{\mathbb{R}^3} f(\boldsymbol{x}')g(\boldsymbol{x}-\boldsymbol{x}') \,\mathrm{d}\boldsymbol{x}'$, which is especially useful for the last two terms in Eq. (2.31).

2.1.3.2 The Kohn–Sham Density Functional Theory

Considering a fluid flow around a cylinder, we might ask ourselves the question[12]: Why should we care about the exact positions and individual movement of all particles rather than describing the flow using simpler, still meaningful, quantities like density, velocity, and energy? The same question can be asked for the electronic structure problem and the individual electrons. And we already have a descriptive quantity at hand, the electronic density

$$\rho_\psi(\boldsymbol{x}) = N \int_{\mathbb{R}^{3(N-1)}} |\psi(\boldsymbol{x}, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N)|^2 \, \mathrm{d}\boldsymbol{x}_2 \cdots \mathrm{d}\boldsymbol{x}_n. \tag{2.33}$$

Density functional theory now tries to formulate the electronic structure problem in terms of ρ_ψ. Indeed, since every ψ_e has a corresponding density ρ_{ψ_e}, we can optimize overall densities that result from a wave function. The idea of representing the ground state in terms of the density goes back[13] to Thomas [252] and Fermi [111], while the development[14] of the density functional theory is attributed to the works of Hohenberg and Kohn [153], Levy [187], and Lieb [188]. Indeed, the ground state minimization problem (2.22) can be reformulated to

$$E_0 = \inf \left\{ E(\rho) + \int_{\mathbb{R}^3} V_\mathrm{nuc}(\boldsymbol{x}) \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \ \rho \in \mathcal{I}_N \right\}, \tag{2.34}$$

with $V_\mathrm{nuc}(\boldsymbol{x})$ from Page 14 defined through the nuclei coordinates $\bar{\boldsymbol{x}}$. The convex set of *N-representable densities* [59] $\mathcal{I}_N$ is characterized by

$$\mathcal{I}_N = \left\{ \rho(\boldsymbol{x}) \geq 0 \mid \sqrt{\rho} \in H^1(\mathbb{R}^3), \ \int_{\mathbb{R}^3} \rho(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = N \right\}. \tag{2.35}$$

The functional $E(\rho)$ in Eq. (2.34) is called *universal*[15] since it is independent of the system under consideration since only the nuclei coordinates, $\bar{\boldsymbol{x}}$, are used to define the potential $V_\mathrm{nuc}(\boldsymbol{x})$ in Eq. (2.34). It reads

$$E(\rho) = \inf \left\{ \langle \psi_\mathrm{e}, (H_\mathrm{e} - V_\mathrm{nuc}) \psi_\mathrm{e} \rangle, \ \psi_\mathrm{e} \in \mathcal{H}_\mathrm{e}, \ \rho_{\psi_\mathrm{e}} = \rho \right\}, \tag{2.36}$$

which shows the connection to the initial ground state minimization in ψ_e when plugging the Eq. (2.36) back into Eq. (2.34) [59]. From a computational point of view, the set $\mathcal{I}_N$ (2.35) looks very tractable: it contains functions of the spatial coordinates in $\mathbb{R}^3$, and we might treat $u := \sqrt{\rho}$ with the constraints $u^2 = \rho \geq 0, \|u\|^2 = N$ and start discretizing the problem. If then, by some miracle[16], $E(u)$ happens to be given by $\int_{\mathbb{R}^3} |\nabla u|^2 \, \mathrm{d}\boldsymbol{x}$, this would not even be that difficult, see, e.g., Brezzi and Fortin [53].

[12] This is a reference to the method of moments used within, e.g., the *kinetic theory of gases* [246].

[13] See [188] and note that the reference [111] is the original publication but, by the time of writing, only a translated version entitled *Statistical method to determine some properties of atoms* seems to be publicly available.

[14] Also, see the prologue *"Early Days of Modern DFT (1964–1979)"* in [60].

[15] It is also called the *Levy–Lieb density functional*, c.f., [59].

[16] This notion is not even entirely implausible; see the first term of the *Thomas–Fermi–Weizsäcker* model in [59], which is an example of an *orbital-free* DFT model [181].

However, the exact form of $E(\rho)$ in Eq. (2.36) is unknown. Thus, we end with an equation with only some explicitly known terms. In contrast, the unknown terms are subject to further approximation or modeling – a situation that, in some other contexts[17], might be called a *closure problem*. A typical approach in such situations is to examine a reference state to extract physical intuition or quantitative information to model the remaining terms.

In the Kohn–Sham model, the system of N non-interacting electrons acts as the reference state, similar to our heuristic motivation at the beginning of Section 2.1.3.1. Under convenient [181] assumptions, the kinetic energy of this reference system reads

$$T_{\mathrm{KS}}(\rho) = \inf\left\{ \frac{1}{2}\sum_{i=1}^{N} \int_{\mathbb{R}^3} |\phi_i|^2 \,\mathrm{d}\boldsymbol{x} \;\middle|\; \phi_i \in H^1(\mathbb{R}^2),\ \int_{\mathbb{R}^3} \phi_i\phi_j \,\mathrm{d}\boldsymbol{x} = \delta_{ij},\ \sum_{i=1}^{N} |\phi_i|^2 = \rho \right\}. \tag{2.37}$$

Defining also the Coulomb energy [59] as

$$J(\rho) = \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\boldsymbol{x})\rho(\boldsymbol{x}')}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}', \tag{2.38}$$

allows the collection of all modeling errors, so far, in the *exchange-correlation functional* (or energy), which is the difference to the true $E(\rho)$ from Eq. (2.36) and is thus defined as

$$E_{\mathrm{xc}}(\rho) = E(\rho) - T_{\mathrm{KS}}(\rho) - J(\rho). \tag{2.39}$$

The abstract problem (2.34) can then be reformulated [59] as

$$\begin{aligned} E_0^{\mathrm{KS}} = \inf\Bigg\{ &\frac{1}{2}\sum_{i=1}^{N} \int_{\mathbb{R}^3} |\phi_i|^2 \,\mathrm{d}\boldsymbol{x} + \int_{\mathbb{R}^3} V_{\mathrm{nuc}}(\boldsymbol{x})\rho(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} + \frac{1}{2}\int_{\mathbb{R}^3}\int_{\mathbb{R}^3} \frac{\rho(\boldsymbol{x})\rho(\boldsymbol{x}')}{\|\boldsymbol{x}-\boldsymbol{x}'\|_2} \,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{x}' \\ &+ E_{\mathrm{xc}}(\rho) \;\middle|\; \phi_i \in H^1(\mathbb{R}^3),\ \int_{\mathbb{R}^3} \phi_i(\boldsymbol{x})\phi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x} = \delta_{ij} \Bigg\}, \end{aligned} \tag{2.40}$$

using $\rho(\boldsymbol{x}) = \sum_{i=1}^{N} |\phi_i(\boldsymbol{x})|^2$ while assuming[18] that Eq. (2.37) holds for some minimizer of Eq. (2.34).

Abstractly speaking, the lack of knowledge of $E(\rho)$ is now transferred to the exchange-correlation functional $E_{\mathrm{xc}}(\rho)$, and approximation methods are needed to model it. Typical approximations [59] include the *local density approximation* (LDA) with $E_{\mathrm{xc}}^{\mathrm{LDA}}(\rho) = \int_{\mathbb{R}^3} \epsilon_{\mathrm{xc}}(\rho) \,\mathrm{d}\boldsymbol{x}$ using, e.g., $\epsilon_{\mathrm{xc}}(\rho) = -C_{\mathrm{D}}\rho^{4/3}$ with the Dirac term $C_{\mathrm{D}} = (3/4)(3/\pi)^{1/3}$, see, e.g., [181], or more sophisticated approximations like the *generalized gradient approximation* (GGA) with $E_{\mathrm{xc}}^{\mathrm{GGA}}(\rho) = \int_{\mathbb{R}^3} f(\rho, \nabla\rho) \,\mathrm{d}\boldsymbol{x}$ for some function f. Note that these approximations can also incorporate empirical information or information from more accurate reference calculations. Given the assumption of differentiability for E_{xc}, the *Kohn–Sham equations*, as described in [181], correspond to the unitary-transformed Euler–Lagrange equations of the minimization

[17] See Footnote 12.

[18] This assumption is not always valid, thus also being part of the Kohn–Sham approximation [59].

problem in (2.40). These equations search for an N-tuple of *Kohn–Sham orbitals* $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_N), \phi_i \in H^1(\mathbb{R}^3)$, such that

$$\begin{cases} K_{\boldsymbol{\rho}_{\boldsymbol{\phi}}} \phi_i(\boldsymbol{x}) = \varepsilon_i \phi_i(\boldsymbol{x}) \\ \langle \phi_i(\boldsymbol{x}), \phi_j(\boldsymbol{x}) \rangle = \delta_{ij} \end{cases}, \tag{2.41}$$

with the *Kohn–Sham operator* [59] given by

$$\varphi(\boldsymbol{x}) \mapsto K_{\boldsymbol{\rho}} \varphi(\boldsymbol{x}) := -\frac{1}{2} \Delta_{\boldsymbol{x}} \varphi(\boldsymbol{x}) + V_{\mathrm{eff},\rho}(\boldsymbol{x}) \varphi(\boldsymbol{x}). \tag{2.42}$$

where, with the *exchange-correlation potential* $V_{\mathrm{xc}}(\boldsymbol{x}) = (\mathrm{d}E_{\mathrm{xc}}(\rho)/\mathrm{d}\rho)(\boldsymbol{x})$, the *effective potential* reads

$$V_{\mathrm{eff},\rho}(\boldsymbol{x}) = V_{\mathrm{nuc}}(\boldsymbol{x}) + \int_{\mathbb{R}^3} \frac{\rho(\boldsymbol{x}')}{\|\boldsymbol{x} - \boldsymbol{x}'\|_2} \, \mathrm{d}\boldsymbol{x}' + V_{\mathrm{xc}}(\boldsymbol{x}). \tag{2.43}$$

Let us formally compare the Kohn–Sham equations (2.41) to the Hartree–Fock equations (2.30). We see that they are very similar formally, and both act on the N-tuple of orbitals $\boldsymbol{\phi}$. The kinetic term, the treatment of the nuclear potential, and the Coulomb energy term have the same form, although both models have a different theoretical derivation. However, in contrast to the molecular orbitals in the HF case, whose Slater determinant approximates the trues wave function, the Kohn–Sham orbitals are not necessarily of a physical nature since only their density is, with proper exchange-correlation treatment, close to the exact electronic density. Due to their similarities, both equations have mostly identical algorithmic treatment, as we will see later in Section 2.1.5.

2.1.3.3 The Gross–Pitaevskii Equation

Let us finish this section on models by considering a mathematical simplification of the KS model (2.42). Choosing $N = 1$, $V_{\mathrm{xc}}(\boldsymbol{x}) = 0$ and replacing the Coulomb interaction $V(\boldsymbol{x} - \boldsymbol{x}') = 1/\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ by an extremely short-range model, i.e., $V(\boldsymbol{x} - \boldsymbol{x}')$ replaced by $\beta\delta(\boldsymbol{x} - \boldsymbol{x}')$ with some $\beta \in \mathbb{R}$ in Eq. (2.43), we would obtain the simplified effective potential $\tilde{V}_{\mathrm{eff},\rho}(\boldsymbol{x}) = V_{\mathrm{ext}}(\boldsymbol{x}) + \rho(\boldsymbol{x})$ (with a reinterpretation of V_{nuc} as a general external potential). Expanding the density $\rho(\boldsymbol{x}) = |\psi_1(\boldsymbol{x})|^2$ and inserting into Eq. (2.41) leads to the single nonlinear eigenvalue problem

$$-\frac{1}{2} \Delta_{\boldsymbol{x}} \psi_1(\boldsymbol{x}) + V_{\mathrm{ext}}(\boldsymbol{x}) \psi_1(\boldsymbol{x}) + \beta |\psi_1(\boldsymbol{x})|^2 \psi_1(\boldsymbol{x}) = \varepsilon_1 \psi_1(\boldsymbol{x}), \tag{2.44}$$

with $\|\psi_1\|_{L^2(\mathbb{R}^3)} = 1$. The Eq. (2.44) is the *Gross–Pitaevskii equation* (GPE), used in studying superfluidity and Bose–Einstein condensates (BECs) [133]. Although these above considerations have no physical significance and are purely mathematical, they show some mathematical connection to the KS model. Thus, one also often encounters the GPE as a simpler model to test algorithms' performance for the nonlinear EVP (2.41). In the proper physical picture of BECs, we deal with *bosons*. At a very low

temperature, all particles are in the same quantum state ψ_{BEC}, i.e., are described by the Hartree product wave function

$$\psi(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N) = \prod_{i=1}^{N} \psi_{\mathrm{BEC}}(\boldsymbol{x}_i). \tag{2.45}$$

The Eq. (2.45) is then used as an ansatz for the electronic ground state problem in Eq. (2.22). After further modeling approximations, one obtains the GPE, see [32, 33, 133, 225].

2.1.4 Discretization Methods

Regarding the numerical solution of the electronic structure problem, we first need to discretize the infinite-dimensional problem into a finite-dimensional problem that is computationally tractable. Just like for any other partial differential equation, we have a variety of different techniques available. We can group the most common approximation methods within the computational chemistry context into *grid-based approximations* and *variational space approximations*.

2.1.4.1 Grid-Based Approximations

The grid-based approximation methods are derived by discretizing the differential operators in the electronic structure problem. Reconsider, for example, the integral $\int_{\mathbb{R}^3}(\rho(\boldsymbol{x}')/\|\boldsymbol{x}-\boldsymbol{x}'\|_2)\,\mathrm{d}\boldsymbol{x}'$ from Eq. (2.43). The calculation of that integral to obtain the so-called *Hartree potential* [38] V_{H} amounts to solving a Poisson equation

$$-\Delta V_{\mathrm{H}}(\boldsymbol{x}) = 4\pi\rho(\boldsymbol{x}). \tag{2.46}$$

Restricting to a large domain, $\boldsymbol{x} = (x_1, x_2, x_3) \in \Omega \subset \mathbb{R}^3$, with zero boundary conditions[19], discretized by a uniform cartesian grid, with N points per spacial direction, as $\Omega_h = \{x_{i,j,k}\}$ with the grid coordinates $x_{i,j,k} := (ih, jh, kh)$, $h := 1/(N+1)$ allows to replace the function $V_{\mathrm{H}} : \Omega \to \mathbb{R}$ by a grid function [29] $V_{\mathrm{H},h} : \Omega_h \to \mathbb{R}$. Then, the partial derivatives can be approximated by using truncated Taylor series [134], e.g., by the second-order central finite difference of the grid function

$$\frac{\partial^2 V_{\mathrm{H}}(\boldsymbol{x})}{\partial x_1^2} \approx h^{-2}(V_{\mathrm{H},h}(x_1-h, x_2, x_3) - 2V_{\mathrm{H},h}(x_1, x_2, x_3) + V_{\mathrm{H},h}(x_1+h, x_2, x_3)). \tag{2.47}$$

Repeating this process and collecting the function values of $V_{\mathrm{H},h}$ in a vector $\boldsymbol{v}_h \in \mathbb{R}^{N^3}$ allows us to write the Poisson equation as a linear system $\boldsymbol{A}\boldsymbol{v}_h = \boldsymbol{\rho}_h$ with the grid point evaluations of $\rho(\boldsymbol{x})$ collected into the vector $\boldsymbol{\rho}_h$. The system has a very structured compact representation with $\boldsymbol{A} = \boldsymbol{I}_N \otimes \boldsymbol{I}_N \otimes \boldsymbol{L}_h + \boldsymbol{I}_N \otimes \boldsymbol{L}_h \otimes \boldsymbol{I}_N + \boldsymbol{L}_h \otimes \boldsymbol{I}_N \otimes \boldsymbol{I}_N$ where $\boldsymbol{L}_h$ represents the second derivative in one spacial direction and $\boldsymbol{I}_N$ the N-identity matrix, see, e.g., [29]. The linear system can then be solved using standard linear

[19] Although real applications require more advanced treatment of boundary conditions, c.f. [181].

algebra techniques. Such a finite difference approach is used for electronic structure calculations (see [20, 37, 38, 108, 109]). Evaluation of integrals by summation rules and further model approximation, such as smooth *pseudopotentials*, often complements these approaches. Based on the nature of the approach, the FD method is also often referred to as a *real-space* method [107].

2.1.4.2 Variational Approaches

In contrast, the variational space approximation methods are based on discretizing the test space of the orbitals ϕ_i. Writing each orbital as a linear combination of basis functions $\phi_i = \sum_{j=1}^{N_\mathrm{b}} c_{i,j}\chi_j$ with N_b basis functions χ_j, and we can discretize the problem by choosing a finite-dimensional subspace $V_\mathrm{b} \subset H^1(\mathbb{R}^3)$ spanned by the basis functions χ_j.

In contrast to the periodic setting within solid-state physics, the basis functions are often centered around the nuclei for the molecular case with a single molecule or structure having N electrons (e.g., in biological applications). This basis of *atomic orbitals* (AO) is then used to construct the *linear combination of atomic orbitals* (LCAO) [59]. For M nuclei with coordinates $\bar{\boldsymbol{x}} = (\bar{\boldsymbol{x}}_1, \dots, \bar{\boldsymbol{x}}_M)$, for a single orbital, this ansatz reads

$$\phi_i(\boldsymbol{x}) = \sum_{j=1}^{M} \sum_{k=1}^{N_\mathrm{b}} c_{i,j,k}\xi_{j,k}(\boldsymbol{x} - \bar{\boldsymbol{x}}_j). \tag{2.48}$$

For the atom-centered basis functions $\xi_{j,k}$, common choices [59, 105] are the *Slater-type orbitals* (STOs), *Gaussian-type orbitals* (GTOs), or general *numerical atomic orbitals* (NAOs). STOs are motivated by the eigenfunctions of the hydrogen atom, i.e., eigenfunctions of linear Schrödinger operators like Eq. (2.2), simplified, up to normalization, and in spherical coordinates, given by $\xi_\mathrm{STO}(r,\theta,\varphi) = r^l e^{-\alpha r} Y_l^m(\theta,\varphi)$ [59] for some varying parameter α. The idea to use GTOs led to significant improvement in evaluating integrals [181], since Gaussian functions allow for a coordinate splitting and dimension reduction when integrated, and are given by $\xi_\mathrm{GTO}(r,\theta,\varphi) = r^l e^{-\beta r^2} Y_l^m(\theta,\varphi)$. Since GTOs cannot represent *nuclear cusps*, an essential feature in electronic structure calculations, combining some STOs and GTOs is also common [59]. Also, the parameters α and β might be optimized, and in general, there is much flexibility for the basis choice since even the complete orbitals can be optimized in the context of NAOs, where the basis is only given numerically [59]. The choice of the correct basis sets seems to be a very active research field, and there are even databases to store and exchange them, e.g., [222]. In the non-molecular setting, popular approaches are based on *plane waves* for the periodic case within, e.g., crystals. Further strategies include, e.g., other spectral- [179], wavelet- [125] or tensor methods [29].

Another discretization method is the *finite element method* (FEM), which usually allows for a much more systematic improvement of the basis. Reconsidering the Poisson problem (2.46) (dropping the 4π-constant and labeling the solution u instead of V_H), we can first write it as a variational problem: Find $u \in V := H_0^1(\Omega)$ such that

$$\forall v \in H_0^1(\Omega): \quad a(u,v) = F(v), \tag{2.49}$$

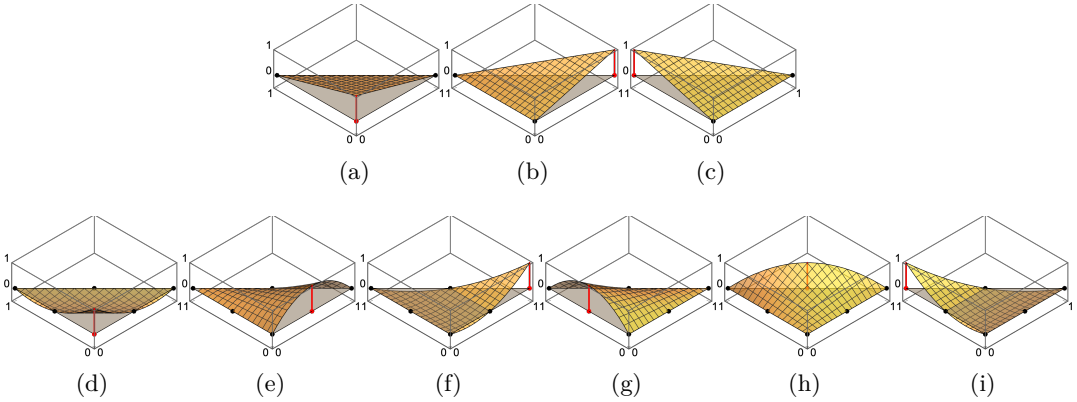


Figure 2.1: Visualization of Lagrangian shape functions for **(a-c)** first-order $\mathbb{P}_1$ and **(d-i)** second-order $\mathbb{P}_2$ elements. Parts of the figure have been used in [246], while the generation script for arbitrary order visualizations can be found in [245].

with the continuous, coercive, bilinear form $a(u, v) = \int_{\mathbb{R}^3} \nabla u \cdot \nabla v \, \mathrm{d}\boldsymbol{x}$, a continuous linear functional $F \in V'$ and $H_0^1(\Omega)$ denoting the Sobolev space of first order having a zero trace. The variational nature can be easily seen since the problem (2.49) characterizes [53] the solution of the minimization problem given by

$$\inf_{v \in H_0^1(\Omega)} \left\{ \frac{1}{2} \int_\Omega |\nabla v|^2 \, \mathrm{d}\boldsymbol{x} - \int_\Omega f v \, \mathrm{d}\boldsymbol{x} \right\}. \tag{2.50}$$

Furthermore, the Lax–Milgram theorem [51] guarantees the existence and uniqueness of Eq. (2.49). The main idea is now to replace V with a finite-dimensional subspace $V_h \subset V$, spanned by the basis functions $\{\chi_j\}_{j=1}^{N_\mathrm{b}}$, and to solve the finite-dimensional problem: Find $u_h \in V_h$ such that

$$\forall v_h \in V_h : \quad a(u_h, v_h) = F(v_h). \tag{2.51}$$

The energy error made by this approximation is quasi-optimal since up to a multiplicative constant, it is optimal, i.e., $\|u - u_h\|_V \leq (C/\alpha) \min_{v \in V_h} \|u - v\|_V$ with the continuity constant C and the coercivity constant α. This result is known as Céa's lemma [51].

Writing the solution as a linear combination of basis functions, i.e., $u_h = \sum_{j=1}^{N_\mathrm{b}} c_j \chi_j$, inserting it into Eq. (2.51), and using the linearity of the bilinear form a and the linear functional F allows us to write the problem as a linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$. The basis coefficients are collected in a vector $\boldsymbol{x} \in \mathbb{R}^{N_\mathrm{b}}$ while the entries of the symmetric *stiffness matrix* $\boldsymbol{A} \in \mathbb{R}^{N_\mathrm{b} \times N_\mathrm{b}}$ are given by $A_{i,j} = a(\chi_i, \chi_j)$ and the right-hand side vector $\boldsymbol{b} \in \mathbb{R}^{N_\mathrm{b}}$ is constructed through the entries $b_i = F(\chi_i)$.

The numerical values within the linear system are determined by fixing the basis, usually done with a geometrical triangulation of the domain Ω. Let $\mathcal{T}_h$ be a conforming

and shape-regular partition of the domain Ω into finite elements $\tau \in \mathcal{T}_h$, where $h := \max_{\tau \in \mathcal{T}_h} \operatorname{diam} \tau$. We can then define the finite element subspace $V_h(\Omega) \subset H_0^1(\Omega)$ as the set of piecewise polynomials with total degree r from the space of polynomials $\mathcal{P}_r$ that satisfy $u|_\tau \in \mathcal{P}_r$ for all $\tau \in \mathcal{T}_h$, such that $|V_h(\Omega)| = N_\mathrm{b}$. The transformation rule allows the integrals to be evaluated using numerical quadrature rules performed on a reference element. For the case of Lagrangian elements in the two-dimensional setting, Fig. 2.1 presents the shape functions, i.e., restriction of the basis functions to one reference element, the 2-simplex for the given case.

Various finite element libraries exist in multiple programming languages, e.g., `FEniCS` [12], `deal.II` [26], `FreeFEM++` [138], `DUNE` [35], `Gridap.jl` [30], to list at least some of them. Usually, most libraries have a generic API, allowing for quick prototyping and easy model adaptation. Changing, e.g., the weak Laplace operator of $-\Delta$ in Eq. (2.49) to $-\Delta + V_\mathrm{ext}(\boldsymbol{x})$, for some regular enough V_ext, and considering the corresponding eigenvalue problem, is thus straightforward and would lead to

$$(\boldsymbol{A} + \boldsymbol{B})\boldsymbol{x} = \lambda \boldsymbol{M}\boldsymbol{x}, \tag{2.52}$$

with the additional symmetric potential matrix $\boldsymbol{B} \in \mathbb{R}^{N_\mathrm{b} \times N_\mathrm{b}}$ and mass matrix $\boldsymbol{M} \in \mathbb{R}^{N_\mathrm{b} \times N_\mathrm{b}}$ with entries $B_{i,j} = \int_\Omega \chi_i(\boldsymbol{x}) V_\mathrm{ext}(\boldsymbol{x}) \chi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$ and $M_{i,j} = \int_\Omega \chi_i(\boldsymbol{x}) \chi_j(\boldsymbol{x}) \,\mathrm{d}\boldsymbol{x}$. Systems of the form (2.52) are essential throughout this thesis since they are the discretized version of the main model problem, the linear Schrödinger equation, see Section 2.1.6.

Coming back to the electronic structure problem, we have to note that the FEM is not a very popular choice within the electronic structure community. A reason is the special meshing requirement, i.e., to resolve certain phenomena near the nuclei [60, 181]. These requirements might be a common drawback of real-space methods but are a general requirement of such grid-based methods [244]. However, works within this category exist, e.g., for the KS equations using adaptivity in [72], using an FE basis only for the radial parts [183, 184], using a DG method for the KS equations [156], and other methods (see, e.g., the reviews [182, 218]). Although inefficient and not widely used in practice, these methods have rigorous mathematical foundations. This fact allows studying phenomena observed within a broader class of numerical methods because they are, e.g., caused by the equations, not their discretization.

2.1.5 Solution Algorithms

Besides discretization, another difficulty in the HF or KS model is the nonlinearity within the set of eigenvalue problems or the energy functionals to minimize. It is possible to directly minimize the functional or iteratively solve the Euler–Lagrange equations, often called *self-consistent field* (SCF) iterations [57]. Although the SCF approach was, some years ago, described as *"the only tractable one"* [181], there is still an active research community dealing with direct minimization methods.

2.1.5.1 Self-Consistent Iterations

For the HF, KS (or GPE model with $N = 1$) from Section 2.1.3.1, after discretization of the orbitals $\phi_i = \sum_{j=1}^{N_\mathrm{b}} c_{i,j} \chi_j$ within a basis, we obtain a set of coupled, nonlinear

EVPs. All orbitals are treated at once with the coefficient matrix $\boldsymbol{C} \in \mathbb{R}^{N_\mathrm{b} \times N}$, leading to a prototypical problem, usually written, c.f., [59, 181] in the form

$$\begin{cases} \boldsymbol{F}(\boldsymbol{D})\boldsymbol{C} = \boldsymbol{S}\boldsymbol{C}\boldsymbol{E} \\ \boldsymbol{C}^T\boldsymbol{S}\boldsymbol{C} = \boldsymbol{I}_N \\ \boldsymbol{D} = \boldsymbol{C}\boldsymbol{C}^T \end{cases}, \tag{2.53}$$

where $\boldsymbol{S} \in \mathbb{R}^{N_\mathrm{b} \times N_\mathrm{b}}$ is the *overlap matrix* (similar to the mass matrix in the FEM case), $\boldsymbol{E} = \operatorname{diag}(\epsilon_1, \ldots, \epsilon_N)$ is the diagonal matrix of lowest eigenvalues of $\boldsymbol{F}(\boldsymbol{D})$ (discretization matrix of either the HF, KS (or GPE) operator), and $\boldsymbol{D} \in \mathbb{R}^{N_\mathrm{b} \times N_\mathrm{b}}$ is the density matrix.

The idea is to start with an initial guess for $\boldsymbol{C}$, build up $\boldsymbol{D}$, and use it to construct an updated $\boldsymbol{F}(\boldsymbol{D})$ such that the linear eigenvalue problem yields an updated $\boldsymbol{C}$, i.e., iterating for $k = 0, 1, \ldots$, the set of equations [59] as

$$\begin{cases} \tilde{\boldsymbol{F}}\boldsymbol{C}_{k+1} = \boldsymbol{S}\boldsymbol{C}_{k+1}\boldsymbol{E}_{k+1} \\ \boldsymbol{C}_{k+1}^T\boldsymbol{S}\boldsymbol{C}_{k+1} = \boldsymbol{I}_N \\ \boldsymbol{D}_{k+1} = \boldsymbol{C}_{k+1}\boldsymbol{C}_{k+1}^T \end{cases}. \tag{2.54}$$

If $\tilde{\boldsymbol{F}} = \boldsymbol{F}(\boldsymbol{D}_k)$, it is the classical SCF approach, called the *Roothaan algorithm* [226] or *pure* SCF [58]. However, convergence and oscillation issues [95] require damping in practice. Choosing $\tilde{\boldsymbol{F}} \mapsto \boldsymbol{F}(\boldsymbol{D}_k) - b\boldsymbol{D}_k$ in Eq. (2.54) yields the *level shifting algorithm* for a shift b. There exists a value $b_0 > 0$ such that the sequence converges to a stationary point with some energy minimization property [58]. When inserted into Eq. (2.54), this strategy effectively changes the energies and thus modifies the local geometry of the energy landscape. We can also use a relaxation and only slightly update the density matrix per step with the mixing rule $\alpha\boldsymbol{D}_{k+1} + (1-\alpha)\boldsymbol{D}_k$. Mixing the operators, e.g., by a linear combination of $\boldsymbol{F}(\boldsymbol{D}_0), \boldsymbol{F}(\boldsymbol{D}_1), \ldots$, under some optimized conditions for the coefficients within this linear combination, is also an option. Such a strategy is often [59] called *direct inversion of the iterative subspace* (DIIS), *Pulay mixing*, *Anderson acceleration*, or *nonlinear GMRES* (see, e.g., [254] for a discussion). Only some of the last steps can be used for the mixing to improve performance.

2.1.5.2 Direct Minimization

The alternative to the SCF approach is to minimize the energy directly. After discretization, this yields a manifold optimization problem since we must incorporate the constraints. Formulations are either based on minimizing $E(\boldsymbol{C})$ on the Stiefel manifold or $E(\boldsymbol{D})$ on the Grassmann manifold (see, e.g., [106] for a review about minimization on manifolds).

Respecting the constraints in combination with the nonlinearity is challenging. The approaches typically use or combine techniques from the optimization community and use, e.g., gradient descent [19, 186], gradient flows [13, 139, 146], Newton [15, 55], and trust-region [113, 154], to name a few.

2.1.6 A Model Problem: The Linear Schrödinger Equation

After reviewing the bigger picture within the electronic structure problem, we now focus on a particular aspect, namely the linear Schrödinger eigenvalue problem: Find (λ, u) such that

$$\begin{cases} -\Delta u + Vu = \lambda u & \text{in } \Omega \\ u = 0 & \text{on } \partial\Omega \end{cases}, \tag{2.55}$$

where V is an external or effective potential. Problems of this kind arise in various applications. If we recall the bigger context and all modeling assumptions from the previous sections, the problem Eq. (2.55) might be placed at the end of the path:

time-dependent Schrödinger equation (2.1) → time-independent Schrödinger equation (2.8) → electronic Hamiltonian with Born–Oppenheimer approximation in atomic units (2.15) → electronic ground state problem (2.22) → HF/KS/GPE equations (2.30), (2.41) and (2.44) → linear EVPs in one SCF cycle (2.54).

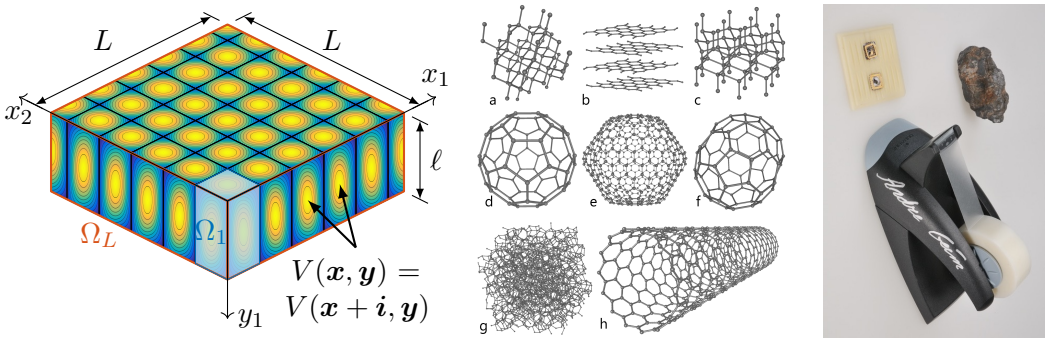
For the Eq. (2.55), we mainly have two options to modify the model: the potential V or the domain Ω. We will discuss both options in the following.

For the operator and, thus, the potential V, we assume periodicity to make the problem interesting and challenging but still mathematically tractable, as we will see in Chapters 3 and 4. Note that we also focus on the case of a single eigenfunction, thus $N = 1$, and given the previous sections, the potential can also be seen as an effective potential V_{eff} similar to (2.43) within the KS operator. Therefore, the model problem could also be interpreted as modeling a single electron within the mean-field potential generated by the other particles in the system. Alternatively, another view is the linearization of more complex models around some state of interest, i.e., the ground state. Furthermore, the potential will be usually complemented with an L^∞-assumption, corresponding to, e.g., the use of pseudopotentials. Note that $V = 0$ reduces to the Laplace eigenvalue problem, which we will frequently use to motivate and highlight difficulties since it has a known set of eigenvalues.

For the domain Ω, we focus on *anisotropy*. In the general case within chemical applications [219], this is differentiated into *chemical anisotropy* based on, e.g., the alignment of atoms, *anisotropy in properties* when specific directions are preferred, e.g., for magnetic properties, and the *domain anisotropy* for domains with very different length scales for the spatial dimensions. We focus on the latter, anisotropic domains with some dimensions expanding, possibly, uniformly while the rest of the dimensions remain fixed. Mathematically, this is described by an abstract d-dimensional box

$$\boldsymbol{z} \in \Omega_L = (0, L)^p \times (0, \ell)^q =: \Omega_{\boldsymbol{x}} \times \Omega_{\boldsymbol{y}} \subset \mathbb{R}^d \quad \text{with } L, \ell \in \mathbb{R}, \tag{2.56}$$

where $\ell = \text{const}$ and $L \to \infty$. Such a domain with a periodic potential V is sketched in Fig. 2.2. These domains were initially motivated by applying the domain decomposition method (see the Section 2.3). However, they are also physically significant, especially for nanomaterials [229]. Think about the Nobel-prize-winning works on graphene [126, 213, 214], which belongs to the class of *2d* material since it



(a) Geometrical setup of Ω_L (Chapter 4). (b) Carbon allotropes [263]. (c) Graphit [264].

Figure 2.2: Visualizations of **(a)** our geometrical model domain Ω_L for $p = 2$ expanding dimensions with $L \to \infty$ from Chapter 4, **(b)** the allotropes of carbon [263], including the anisotropic cases of graphene and nanotubes, and **(c)** a representation of the Nobel-prize-winning experiment of graphene extraction with a tape dispenser and graphite, located in the Nobel museum in Stockholm [264].

is composed of the regular alignment of carbon atoms in a hexagonal lattice, with the thickness of only one atom. The material is, thus, anisotropic. Initially extracted with only a tape dispenser (Fig. 2.2c) from the three-dimensional structure graphite, its unique geometry implies various exciting material properties. Further applications include *1d* structures like carbon nanotubes, carbon nanowires, or *2.5d* structures created by stacking two twisted plane materials; see Fig. 2.2b for the visualization of carbon allotropes (also see Section 4.1.3 for a further discussion).

The boundary conditions are again motivated by the DD inspiration for the case of linear systems, but for $L \to \infty$, a periodic V allows a unique perspective on our model with exciting observations: the model is located *in between* the theory of fully periodic systems and the fully static case. This fact makes the eigenvalue equation, combined with the domain assumptions, a rather nonstandard problem with some unique features[20]. In fact, for $L \to \infty$, the difference between the first and second eigenvalue can become arbitrarily small, which poses significant challenges for the solution to the discretized Eq. (2.52).

2.2 Iterative Eigenvalue Algorithms

We will now move to the iterative algorithms to solve the discretized eigenvalue problems of the type (2.52), i.e.

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{B}\boldsymbol{x}, \tag{2.57}$$

[20] For example, the finite-dimensional factorization of Theorem 3.1 could, e.g., be related to the infinite-dimensional periodic case (*Bloch theorem*).

with the symmetric, positive-definite matrices $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$, the eigenvector $\boldsymbol{x} \in \mathbb{R}^n$, and the eigenvalue $\lambda \in \mathbb{R}$. Depending on the discretization, $\boldsymbol{B}$ is not necessarily the identity matrix, and the Eq. (2.57) is thus called a *generalized eigenvalue problem* (GEVP). For simplicity of the presentation, we set $\boldsymbol{B} = \boldsymbol{I}_n$. Note that this can easily be archived by transforming the problem to a standard EVP, i.e., $\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{B}\boldsymbol{x} \Leftrightarrow \boldsymbol{B}^{-1}\boldsymbol{A}\boldsymbol{x} = \lambda\boldsymbol{x}$. For actual numerical implementation, we will not use this transformation since it is beneficial to work in the changed $\boldsymbol{B}$-metric, using $\|\cdot\|_{\boldsymbol{B}}$ as the $\boldsymbol{B}$-norm, i.e., $\|\boldsymbol{x}\|_{\boldsymbol{B}} = \sqrt{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}}$ when normalization is applied and using $\boldsymbol{A} - \sigma\boldsymbol{B}$ when shifting of the spectrum is required. We will also assume that the eigenvalues are ordered, i.e., $0 \leq \lambda^{(1)} \leq \lambda^{(2)} \leq \cdots \leq \lambda^{(n)}$ with corresponding eigenvectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$.

Since we deal with the discretization of PDE operators on a localized basis, we only consider the iterative approach. The algebraic methods are usually infeasible due to memory limitations since the resulting matrices are large and sparse. Due to the previously presented application cases, we are only interested in the extremal eigenvalues (the smallest eigenvalues for the ground state calculations). However, the method generalizes the interior eigenpairs when properly shifting the matrix (since interior eigenvalues can become extremal ones). For a more detailed overview, we refer, e.g., to the classical books [217, 266], the more numerical works [31, 227, 256], the review article [128] and references therein, or the corresponding sections within the books [71, 96, 129, 155, 255, 267].

2.2.1 Power Method

We start with the *power method* (PM), attributed[21] to Müntz [204, 205] and von Mises[22] [259], that repeatedly applies the matrix $\boldsymbol{A}$ to a vector $\boldsymbol{x}_0$ and normalizes the result subsequently, i.e.,

$$\frac{\boldsymbol{x}_0}{\|\boldsymbol{x}_0\|_2}, \frac{\boldsymbol{A}\boldsymbol{x}_0}{\|\boldsymbol{A}\boldsymbol{x}_0\|_2}, \frac{\boldsymbol{A}^2\boldsymbol{x}_0}{\|\boldsymbol{A}^2\boldsymbol{x}_0\|_2}, \ldots \tag{2.58}$$

The sequence (2.58) converges to the eigenvector corresponding to the largest eigenvalue of $\boldsymbol{A}$ if the initial vector $\boldsymbol{x}_0$ is chosen appropriately and the largest eigenvalue is simple. The Algorithm 1 (with $\alpha = 1, \sigma_k = 0$) presents the idea in a structured way. For a given approximation to the eigenvector, we can use the *Rayleigh quotient*, $R_{\boldsymbol{A}}(\boldsymbol{x}) := (\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})/(\boldsymbol{x}^T\boldsymbol{x})$, to get a corresponding eigenvalue approximation with $\lambda_k = R_{\boldsymbol{A}}(\boldsymbol{x}_k)$.

To understand the convergence of the algorithm, we expand the initial vector in the eigenvectors of $\boldsymbol{A}$, i.e., $\boldsymbol{x}_0 = \sum_{i=1}^n c_i\boldsymbol{x}^{(i)}$, with the eigenvalues $\lambda^{(1)} \leq \cdots < \lambda^{(n)}$ and the corresponding eigenvectors $\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(n)}$. With the constants d_k corresponding

[21] See [128].
[22] Thus also known as the *von Mises iteration* in some contexts.

to normalization, the sequence (2.58) becomes [255]

$$\boldsymbol{x}_k = d_k(\boldsymbol{A}^k\boldsymbol{x}_0) = d_k\left(\sum_{i=1}^{n} c_i(\lambda^{(i)})^k\boldsymbol{x}^{(i)}\right) = d_k(\lambda^{(n)})^k\left(\sum_{i=i}^{n-1} c_i\left(\frac{\lambda^{(i)}}{\lambda^{(n)}}\right)^k\boldsymbol{x}^{(i)} + c_n\boldsymbol{x}^{(n)}\right). \tag{2.59}$$

In Eq. (2.59), we observe the contraction of components in all but one direction due to $(\lambda^{(i)}/\lambda^{(n)})$-terms. More precisely, we have the following.

Theorem 2.1 (Convergence of the power method [255])**.** *Suppose* $|\lambda^{(1)}| \leq \cdots \leq |\lambda^{(n-1)}| < |\lambda^{(n)}|$ *and* $\boldsymbol{x}_0^T\boldsymbol{x}^{(n)} \neq 0$ *(i.e.,* $\boldsymbol{x}_0$ *has some component in* $\boldsymbol{x}^{(n)}$*-direction). Then, the iterates of the Algorithm 1 satisfy*

$$\|\boldsymbol{x}_k - (\pm\boldsymbol{x}^{(1)})\| \in \mathcal{O}(\rho(\boldsymbol{A})^k), \quad |\lambda_k - \lambda^{(1)}| \in \mathcal{O}(\rho(\boldsymbol{A})^{2k}), \tag{2.60}$$

as $k \to \infty$*, with* $\rho(\boldsymbol{A}) = |\lambda^{(n-1)}/\lambda^{(n)}|$ *denoting the convergence rate. The* $\pm$ *sign means that one or the other choice of sign is to be taken at each step* k*.*

Thus, the distribution of eigenvalue and especially the ratio between the desired and the first closest neighboring eigenvalue determines the convergence speed. In practice, typical convergence criteria in Algorithm 1 are based on the *spectral residual* $\boldsymbol{r}(\boldsymbol{x}_k) := \boldsymbol{A}\boldsymbol{x}_k - R_{\boldsymbol{A}}(\boldsymbol{x}_k)\boldsymbol{x}_k$, e.g., when $\|\boldsymbol{r}_k\|_2 \leq \texttt{TOL}$ is reached. The spectral residual will play an essential role for gradient-based methods in Section 2.2.5 (since it points in the direction of the Rayleigh quotient gradient), and it can be easily seen that $\boldsymbol{r}(\boldsymbol{y}) = \boldsymbol{0}$ corresponds to $\boldsymbol{y}$ being an eigenvector. The algorithm's advantage is its simplicity since only matrix-vector multiplications are performed. If the lowest eigenvalue, see the Section 2.1, is of interest, the algorithm must be modified.

Algorithm 1 Generic single vector iteration: power method ($\alpha = 1, \sigma_k = 0$), inverse power method ($\alpha = -1, \sigma_k = 0$), shifted inverse power method ($\alpha = -1, \sigma_k = \sigma$), and Rayleigh quotient iteration ($\alpha = -1, \sigma_k = \lambda_k$)

Require: a matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, an initial vector $\boldsymbol{x}_0 \in \mathbb{R}^n$

Normalize $\boldsymbol{x}_0 := \boldsymbol{x}_0/\|\boldsymbol{x}_0\|_2$
Initialize $k := 0$
while not converged **do**
    Calculate $\boldsymbol{x}_{k+1} := (\boldsymbol{A} - \sigma_k\boldsymbol{I}_n)^\alpha\boldsymbol{x}_k$
    Normalize $\boldsymbol{x}_{k+1} := \boldsymbol{x}_{k+1}/\|\boldsymbol{x}_{k+1}\|_2$
    Calculate $\lambda_{k+1} := R_{\boldsymbol{A}}(\boldsymbol{x}_{k+1})$
    Increment $k := k + 1$
end while
return eigenpair approximation $(\lambda_k, \boldsymbol{x}_k)$

2.2.2 Inverse Power Method

In our applications, we are typically interested in the lowest eigenpair. Thus, we use the *inverse power method*, which replaces the matrix $\boldsymbol{A}$ by $\boldsymbol{A}^{-1}$ in the line 4 of Algorithm 1. The resulting method converges to the smallest eigenpair with appropriate $\boldsymbol{x}_0$ since the eigenvalues essentially change their ordering with $\lambda^{(i)}(\boldsymbol{A}^{-1}) = 1/\lambda^{(i)}(\boldsymbol{A})$ (in our positive definite case).

The convergence factor in Theorem 2.1, $\rho(\boldsymbol{A}^{-1}) = |\lambda^{(1)}/\lambda^{(2)}|$, is called the *fundamental ratio* $r(\boldsymbol{A})$. It depends on the *fundamental gap*, $g(\boldsymbol{A}) := \lambda^{(2)} - \lambda^{(1)}$, by the relation $r(\boldsymbol{A}) = \lambda^{(1)}/(\lambda^{(1)} + g(\boldsymbol{A}))$. If a matrix $\boldsymbol{A}_\epsilon$ depends on a parameter ϵ (e.g., the mesh size, domain length, the strength of the external potential), the gap might depend on ϵ and vanish in the limit, i.e., $g_\epsilon(\boldsymbol{A}_\epsilon) \to 0$ for $\epsilon \to 0$. This vanishing gap would then result in $r_\epsilon(\boldsymbol{A}) \to 1$ leading to an arbitrary bad convergence rate. This situation happens, e.g., for the model problem from Section 2.1.6 if $L \to \infty$. For such cases, or any other cases with a small gap, a shift $\sigma \in \mathbb{R}$ can be applied since shifting affects only the spectrum but not the eigenfunctions.

Replacing $\boldsymbol{A}$ by $(\boldsymbol{A} - \sigma \boldsymbol{I}_n)^{-1}$ in line 4 of Algorithm 1 leads to the *shifted inverse power method*. The shift-and-invert (SI) strategy shifts the spectrum uniformly by the value σ. The convergence factor then reads $\rho_\sigma(\boldsymbol{A}) = (\lambda^{(1)} - \sigma)/(\lambda^{(2)} - \sigma)$ (again provided that $0 < \sigma < \lambda^{(1)} < \lambda^{(2)} \leq \cdots$). We observe that the closer the σ is to $\lambda^{(1)}$, the smaller the ratio. In the extreme case of $\sigma = \lambda^{(1)}$, we end up with the singular matrix $\boldsymbol{A} - \lambda^{(1)}\boldsymbol{I}_n$. The pseudo-inverse of that matrix, $(\boldsymbol{A} - \lambda^{(1)}\boldsymbol{I}_n)^\dagger$, is an orthogonal projector onto the space spanned by all except the first eigenvector projector and we could use it to remove all unwanted directions of a starting vector $\boldsymbol{x}_0$ to solve the system immediately; this is also the reason, why it is seen as the *perfect preconditioner* [170]. An analogy might be using the preconditioner $\boldsymbol{M} = \boldsymbol{A}^{-1}$ when dealing with linear systems, which is as hard to apply as solving the system. For the eigenvalue case, $\lambda^{(1)}$ is, of course, unknown and part of the solution.

The inverse power method is more expensive than the power method since it now involves the solution to a linear system per step (or an expensive factorization a-priori). This direct solution strategy is later used in Chapter 3. Alternatively, and since we deal with PDE-based problems, it is thus often beneficial to use iterative solvers for the linear system, which we will discuss in Chapter 4 after constructing a suitable preconditioner. We refer to that strategy as *inner-outer* iterative eigensolvers since it leads to another loop to solve each linear system per step, see [114, 115]. Further, note that the shift-and-invert techniques are not the only possible spectral transformations that can be used. We refer to, e.g., [31] that also discusses the *Cayley transformation* $(\boldsymbol{A} - \sigma\boldsymbol{I}_n)^{-1}(\boldsymbol{A} + \mu\boldsymbol{I}_n)$ by prescribing a root μ and a pole σ. Generally, any rational function in $\boldsymbol{A}$ can be used (see the related *filtering* [27, 268] techniques and the *FEAST* algorithm [221]). Using the current eigenvalue estimate λ_k as a shift, i.e., an adaptive shift, is the essential idea of the following method.

2.2.3 Rayleigh Quotient Iteration

The *Rayleigh quotient iteration* (RQI) updates the shift σ in each iteration and uses the current eigenvalue estimate, i.e. by using the matrix $(\boldsymbol{A} - \lambda_k \boldsymbol{I}_n)^{-1}$. The idea,

usually attributed[23] to Wieland [262], is excellent since we know that the standard inverse iteration has $\lambda_k \to \lambda^{(1)}$ and that the perfect preconditioner matrix is precisely built with $(\boldsymbol{A} - \lambda^{(1)}\boldsymbol{I}_n)$ (see the discussion in the last section). The method has a cubic convergence but may not converge to the smallest eigenpair [31]. It also comes with the increased computational cost for factorization since the matrix changes in each iteration step. Constructing more reliable convergence is active research (see, e.g. [116]). Furthermore, we want to mention the inherent connection of the RQI and Newton's method; see [31] for a detailed review.

2.2.4 Block Iterations

Generalizing all the presented single vector iterations to block iterations searching for multiple, say, p orthonormal eigenvectors simultaneously is straightforward. Instead of a single $\boldsymbol{x} \in \mathbb{R}^n$, a tall matrix $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ is used within the framework of Algorithm 1. For instance, the iteration of the power method would then read as

$$\tilde{\boldsymbol{X}}_{k+1} = \boldsymbol{A}\boldsymbol{X}_k, \tag{2.61}$$

followed by an orthogonalization [255] using the QR factorization, as $\tilde{\boldsymbol{X}}_{k+1} = \boldsymbol{X}_{k+1}\boldsymbol{R}_{k+1}$, or the *(modified) Gram–Schmidt* procedure. Although not the focus for use within our model problem Eq. (2.55), block iterations are essential for applications in electronic structure calculations (see Section 2.1) since the multiple vectors correspond to the system's orbitals.

2.2.5 Gradient-Based Methods

In Section 2.1, an energy minimization problem led to eigenvalue problems, the HF or KS equation. Using the Rayleigh quotient, we can also use an energy minimization approach in the discrete case. The idea is to minimize the Rayleigh quotient, i.e., the energy, concerning the eigenvector $\boldsymbol{x}$ since we know from the *min-max theorem* [227] that $\lambda^{(1)} = \min_{\boldsymbol{x}\in\mathbb{R}^n} R_{\boldsymbol{A}}(\boldsymbol{x})$. Thus, using a *gradient descent* method is natural to minimize the Rayleigh quotient. Using matrix calculus, we obtain the gradient of the Rayleigh quotient as

$$\nabla R_{\boldsymbol{A}}(\boldsymbol{x}) = \frac{2}{\boldsymbol{x}^T\boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x} - R_{\boldsymbol{A}}(\boldsymbol{x})\boldsymbol{x}). \tag{2.62}$$

Having access to that gradient, we can apply all the standard techniques from optimization theory. Furthermore, the gradient Eq. (2.62) points in the same direction as the spectral residual $\boldsymbol{r}_k := \boldsymbol{A}\boldsymbol{x}_k - R_{\boldsymbol{A}}(\boldsymbol{x}_k)\boldsymbol{x}_k$, which is already available in the algorithm since it acts as a convergence measure. Thus, using it also as a descent direction is handy.

2.2.5.1 Steepest Descent Methods

Starting from an initial guess $\boldsymbol{x}_0 \in \mathbb{R}^n$, the steepest descent method, per iteration, goes a step towards the negative gradient (see, e.g., the optimization literature for

[23] See [243] for an excellent discussion about its history; thus also called *Wieland iteration*.

more details [174, 212]). The iteration is then performed using the abstract rule given by

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \tau_k \boldsymbol{P}_k \boldsymbol{d}_k, \tag{2.63}$$

where we set the direction $\boldsymbol{d}_k = -\boldsymbol{r}_k \in \mathbb{R}^n$ (which is proportional to $\nabla R_{\boldsymbol{A}}(\boldsymbol{x}_k)$) and the preconditioner, $\boldsymbol{P} = \boldsymbol{I}_n \in \mathbb{R}^{n\times n}$. If $\boldsymbol{P} \neq \boldsymbol{I}_n$, we call it the *preconditioned steepest descent* method. The step size $\tau_k \in \mathbb{R}$ could be a constant, adaptively modified in each iteration by some rule, or can be chosen based on the usual *line search* algorithms, e.g., the *Armijo* or *Wolfe* conditions [212].

For the present case of eigenvalue problems, using the *locally optimal steepest descent* method, which uses the optimal step size τ_k in each iteration, is beneficial. The optimal step size is the minimizer of the Rayleigh quotient along the direction $\boldsymbol{d}_k$, i.e., defined as

$$\tau_k = \operatorname*{arg\,min}_{\tau\in\mathbb{R}} R_{\boldsymbol{A}}(\boldsymbol{x}_k + \tau \boldsymbol{d}_k). \tag{2.64}$$

Choosing the preconditioner matrix $\boldsymbol{P}_k = \boldsymbol{A}^{-1}$ is sometimes called *preconditioned inverse interation* (PINVIT) [207]. In fact, with $\tau_k = 1$, we obtain for Eq. (2.63) that

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \boldsymbol{A}^{-1}(\boldsymbol{A}\boldsymbol{x}_k - R_{\boldsymbol{A}}(\boldsymbol{x}_k)\boldsymbol{x}_k) = R_{\boldsymbol{A}}(\boldsymbol{x}_k)\boldsymbol{A}^{-1}\boldsymbol{x}_k, \tag{2.65}$$

which is, after normalization, one step of the inverse power method. For more results about this class of methods, we refer, e.g., to [25, 48, 168, 169, 170, 172, 173, 208, 209, 210, 273]. The choice of $\boldsymbol{P}_k = \boldsymbol{A}^{-1}$ is not uncommon as a preconditioner. Usage of this preconditioner seems very expensive from the perspective of iterative linear solvers since we will have to solve the system in each iteration (or perform an expensive factorization). However, for typical PDE-based problems, the methods perform poorly without preconditioning, and the cost of the preconditioner is usually negligible compared to the cost of the eigenvalue problem. However, usually, the strict equality for $\boldsymbol{P}_k$ can be replaced by a spectral equivalence requirement [25], i.e.,

$$(1-\gamma)\boldsymbol{x}^T\boldsymbol{P}_k^{-1}\boldsymbol{x} \leq \boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \leq (1+\gamma)\boldsymbol{x}^T\boldsymbol{P}_k^{-1}\boldsymbol{x}, \quad \forall \boldsymbol{x} \in \mathbb{R}^n \text{ and some } \gamma \in [0,1). \tag{2.66}$$

The solution to the optimization problem (2.64) remains to be discussed. For the present case of the eigenvalue problem, this can be quickly done using the *Rayleigh–Ritz* (RR) procedure, a standard tool for iterative eigenvalue solvers in general, see [31].

2.2.5.2 Rayleigh–Ritz Procedure

We motivate the RR procedure by considering the minimization problem (2.64) and follow the presentation of [31]. Let us assume that we have a subspace $\mathcal{K} \subset \mathbb{R}^n$ with dimension m, and we want to find approximate eigenpairs of $\boldsymbol{A}\tilde{\boldsymbol{x}} = \tilde{\lambda}\tilde{\boldsymbol{x}}$ where $\tilde{\boldsymbol{x}} \in \mathcal{K}$ and $\tilde{\lambda} \in \mathbb{R}$. Of course, if $\mathcal{K}$ does not contain eigenvectors of $\boldsymbol{A}$, then we can not solve the eigenvalue problem within $\mathcal{K}$. However, we can find the best approximation in $\mathcal{K}$ by imposing the *Galerkin condition*, i.e., the *best approximation* in the Euclidian norm, as

$$\boldsymbol{A}\tilde{\boldsymbol{x}} - \tilde{\lambda}\tilde{\boldsymbol{x}} \perp \mathcal{K}. \tag{2.67}$$

To transfer this orthogonality condition into a matrix problem, we define the matrix of basis vectors $\boldsymbol{V} = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m) \in \mathbb{R}^{n \times m}$ where $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_m\}$ is an orthonormal basis of $\mathcal{K}$. After writing the unknown $\tilde{\boldsymbol{x}}$ in that basis as $\tilde{\boldsymbol{x}} = \boldsymbol{V}\boldsymbol{y}$ with $\boldsymbol{y} \in \mathbb{R}^m$, we obtain the matrix problem as

$$\forall i = 1, \ldots, m: \quad \boldsymbol{v}_i^T(\boldsymbol{A}\boldsymbol{V}\boldsymbol{y} - \tilde{\lambda}\boldsymbol{V}\boldsymbol{y}) = \boldsymbol{0}, \tag{2.68}$$

which translates to the m-dimensional eigenvalue

$$\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{y} = \tilde{\lambda}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{y}. \tag{2.69}$$

In practice, we use this method to compute the optimal stepwidth τ_k in Eq. (2.64). Following [24], we can do the calculation and expand $\min_{\tau \in \mathbb{R}} R_{\boldsymbol{A}}(\boldsymbol{x}_k + \tau \boldsymbol{d}_k)$ as

$$\min_{\tau \in \mathbb{R}} \frac{(\boldsymbol{x}_k + \tau \boldsymbol{d}_k)^T \boldsymbol{A}(\boldsymbol{x}_k + \tau \boldsymbol{d}_k)}{(\boldsymbol{x}_k + \tau \boldsymbol{d}_k)^T(\boldsymbol{x}_k + \tau \boldsymbol{d}_k)} = \min_{\tau \in \mathbb{R}} \frac{\begin{pmatrix} 1 \\ \tau \end{pmatrix}^T \begin{pmatrix} \boldsymbol{x}_k^T\boldsymbol{A}\boldsymbol{x}_k & \boldsymbol{x}_k^T\boldsymbol{A}\boldsymbol{d}_k \\ \boldsymbol{d}_k^T\boldsymbol{A}\boldsymbol{x}_k & \boldsymbol{d}_k^T\boldsymbol{A}\boldsymbol{d}_k \end{pmatrix} \begin{pmatrix} 1 \\ \tau \end{pmatrix}}{\begin{pmatrix} 1 \\ \tau \end{pmatrix}^T \begin{pmatrix} \boldsymbol{x}_k^T\boldsymbol{x}_k & \boldsymbol{x}_k^T\boldsymbol{d}_k \\ \boldsymbol{d}_k^T\boldsymbol{x}_k & \boldsymbol{d}_k^T\boldsymbol{d}_k \end{pmatrix} \begin{pmatrix} 1 \\ \tau \end{pmatrix}}, \tag{2.70}$$

where the last expression attains its minimum if the vector $(1, \tau)^T$ points in the same direction as the smallest eigenvector, $\boldsymbol{y}^{(1)} \in \mathbb{R}^2$, of the generalized, two-dimensional eigenvalue problem $\tilde{\boldsymbol{A}}\boldsymbol{y} = \tilde{\lambda}\tilde{\boldsymbol{B}}\boldsymbol{y}$, where $\tilde{\boldsymbol{A}} = \boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}$, $\tilde{\boldsymbol{B}} = \boldsymbol{V}^T\boldsymbol{V}$, with the matrix $\boldsymbol{V} = (\boldsymbol{x}_k, \boldsymbol{d}_k)$. So if $\boldsymbol{y}^{(1)}$ has no trivial first component, i.e., $\boldsymbol{y}_1^{(1)} \neq 0$, then taking $\tau_k = \boldsymbol{y}_2^{(1)}/\boldsymbol{y}_1^{(1)}$ yields the optimal step size in the Eq. (2.64) by the scaling invariance of the Rayleigh quotient. This non-triviality condition is irrelevant for practical calculations[24], but we can also directly minimize in the space spanned by $\boldsymbol{x}_k$ and $\boldsymbol{d}_k$ with

$$\boldsymbol{x}_k = \operatorname*{arg\,min}_{\boldsymbol{y} \in \operatorname{span}\{\boldsymbol{x}_k, \boldsymbol{d}_k\} \setminus \{\boldsymbol{0}\}} R_{\boldsymbol{A}}(\boldsymbol{y}), \tag{2.71}$$

since we need to solve a two-dimensional eigenvalue problem in any case. Geometrically, we replaced the (affine) line search with a (linear) subspace search.

After presenting the Rayleigh–Ritz procedure, we also want to highlight the class of subspace iterations where the search dimension increases. Recall that we calculated the sequence $\boldsymbol{x}_0, \boldsymbol{A}\boldsymbol{x}_0, \boldsymbol{A}^2\boldsymbol{x}_0, \ldots$ for the power method but only used the previous iterate in the next iteration. The mentioned sequence, however, constructs a *Krylov space* defined by $\mathcal{K}_i(\boldsymbol{A}, \boldsymbol{x}_0) = \operatorname{span}\{\boldsymbol{x}_0, \boldsymbol{A}\boldsymbol{x}_0, \ldots, \boldsymbol{A}^i\boldsymbol{x}_0\}$. We can use the Rayleigh–Ritz procedure to approximate the eigenpairs within the current $\mathcal{K}_k(\boldsymbol{A}, \boldsymbol{x}_0)$. Popular Krylov space algorithms are the *Arnoldi* method for general and the *Lanczos* method for Hermitian matrices [227]. At the same time, other related methods like the *Jacobi-Davidson* method also consider subspaces with increasing dimensions, see [24, 227].

[24] Thus, it is usually not discussed.

2.2.5.3 Locally Optimal Preconditioned Conjugated Gradients Method

Based on the optimal choice within a two-dimensional space (Eq. (2.71)), we can also optimize in the space spanned by the current iterate $\boldsymbol{x}_k$, the previous iterate $\boldsymbol{x}_{k-1}$, and the preconditioned residual $\boldsymbol{P}_k\boldsymbol{r}_k$. This choice then leads to a three-term recurrence and defines the so-called *locally optimal preconditioned conjugated gradient* (LOPCG) method [167]

$$S_k = \operatorname{span}\{\boldsymbol{x}_k, \boldsymbol{x}_{k-1}, \boldsymbol{P}_k\boldsymbol{r}_k\}, \quad \tilde{\boldsymbol{x}}_{k+1} = \operatorname*{arg\,min}_{\boldsymbol{y}\in S_k\setminus\{\boldsymbol{0}\}} R_{\boldsymbol{A}}(\boldsymbol{y}), \quad \boldsymbol{x}_{k+1} = \tilde{\boldsymbol{x}}_{k+1}/\|\tilde{\boldsymbol{x}}_{k+1}\|_2. \tag{2.72}$$

The terminology of the method was inspired [167, 171] by the conjugated gradient (CG) method for linear systems (which also uses a three-term recurrence and minimizes the energy error per iteration [98]) and is then combined with the local optimality obtained by using the Rayleigh–Ritz method within the subspace S_k. However[25], note that the results from the linear case can not be directly transferred to the eigenvalue case since the Rayleigh quotient is not quadratic (see the discussion in [24, 110] and references therein). This method performs very well (see the results in Chapter 3) with slightly increased memory requirements and costs due to the three-dimensional optimization search compared to the locally optimal steepest descent variant.

The method also needs a suitable preconditioner, usually taken as $\boldsymbol{A}^{-1}$ or some approximation. We link to the discussion in Chapter 4 when constructing a suitable preconditioner based on the domain decomposition method and apply it in an inner-outer scheme.

2.3 Domain Decomposition Methods

This section introduces the related concept of the domain decomposition (DD) method. Although the initial idea of the method is quite old, the method has been developed and improved over the last decades and is still an active field of research. Since our main focus will be overlapping Schwarz methods to use as preconditioners for the discrete cases, we can only discuss some of the different flavors, tweaks, and application cases. We thus refer the interested reader to the books by Smith [234], Quarteroni and Valli [223], Toselli and Widlund [253], Mathew [198], Dolean, Jolivet, and Nataf [98], and the corresponding chapter in the book of Ciaramella and Gander [89].

2.3.1 Continuous Domain Decomposition Methods

In the original paper [231] from 1870, Schwarz was interested in the solution to the Laplace equation $\Delta u = 0$ in a domain $\Omega \subset \mathbb{R}^n$ with boundary $\partial\Omega$ and boundary condition $u = g$ on $\partial\Omega$. Due to the lack of theoretical tools [98], see also [122] for a

[25] In [171], it was shown that the method behaves similarly to the classical CG method applied to the singular system $\boldsymbol{A} - \lambda^{(1)}\boldsymbol{I}$, see also [215] for a discussion. Using the CG algorithm for the eigenvalue problem recovers some known properties only in the asymptotic limit, as discussed in [110]. However, this can depend on the specific implementation.

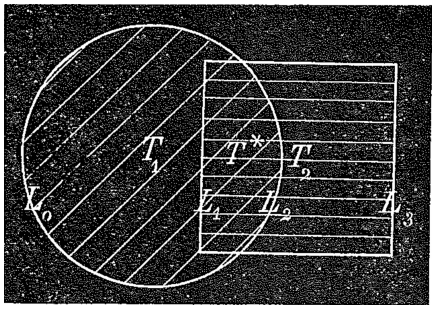


Figure 2.3: The original DD sketch by Schwarz in [231] with two subdomains, T_1 and T_2, that overlap in the region $T^* = T_1 \cap T_2$.

very detailed discussion, he proposed to decompose the domain into problems within simpler domains for which existing tools can be used (see the Fig. 2.3 for the original sketch of this decomposition). He chose two overlapping subdomains, Ω_1 and Ω_2, where $\Omega = \Omega_1 \cup \Omega_2$ and $\Omega_1 \cap \Omega_2 \neq \emptyset$. Then, the solution to the problem on the whole domain is the limit of $k = 1, 2, \ldots$ to the following sequence

$$\begin{aligned} \Delta u_1^k &= 0 \text{ in } \Omega_1 \ (=T_1) & \Delta u_2^k &= 0 \text{ in } \Omega_2 \ (=T_2) \\ u_1^k &= g \text{ on } \partial\Omega \cap \overline{\Omega}_1 \ (=L_0) \ \rightarrow & u_2^k &= g \text{ on } \partial\Omega \cap \overline{\Omega}_2 \ (=L_3). \\ u_1^k &= u_2^{k-1} \text{ on } \Gamma_1 \ (=L_2) & u_2^k &= u_1^n \text{ on } \Gamma_2 \ (=L_1) \end{aligned} \tag{2.73}$$

The sequence (2.73) considers each subdomain separately and uses the outer boundary condition g where possible. The value of the function within the other subdomain is taken for the boundary parts that do not belong to the global outer boundary $\partial\Omega$. Using the maximum principle, the convergence of that sequence can be proven [122]. Since the equations still need to be discretized, we refer to this strategy as a *continuous domain decomposition method*. The method is also called the *Schwarz alternating method*.

Many years later, at a time when parallel computers became more and more available [122], Lions studied the method in a series of papers [190, 191, 192] and proposed in [192, p18] a parallel version of the method (in adapted notation) as

$$\begin{aligned} \Delta u_1^k &= 0 \text{ in } \Omega_1 \ (=T_1) & \Delta u_2^k &= 0 \text{ in } \Omega_2 \ (=T_2) \\ u_1^k &= g \text{ on } \partial\Omega \cap \overline{\Omega}_1 \ (=L_0) \ \rightarrow & u_2^k &= g \text{ on } \partial\Omega \cap \overline{\Omega}_2 \ (=L_3). \\ u_1^k &= u_2^{k-1} \text{ on } \Gamma_1 \ (=L_2) & u_2^k &= u_1^{k-1} \text{ on } \Gamma_2 \ (=L_1) \end{aligned} \tag{2.74}$$

This minor change had a significant impact on parallel computing. The problems on each subdomain are now independent since they only depend on the boundary values from the $(k-1)$-th step. While the former, Eq. (2.73), operates in a *Gauß–Seidel* style, the latter operates in a *Jacobi* style when using the terminology of iterative solvers. The convergence of both methods typically depends on the overlap [98] since more overlap allows for more information exchange between the subdomains intuitively.

$$\begin{pmatrix} a_{11} & a_{12} & 0 & \cdots & 0 \\ a_{21} & a_{22} & a_{23} & \cdots & 0 \\ 0 & a_{32} & a_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_n \end{pmatrix}$$

Figure 2.4: A minimal algebraic non-overlapping decomposition of $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

The *non-overlapping Schwarz method* is the extreme case when the subdomains do not overlap. The Dirichlet boundary condition will never change for such cases, leading to no information exchange between the subdomains. Thus, the boundary conditions are changed to Robin-type transmission conditions [98], e.g., of the form $(\partial_n + \alpha)u_1^k = (\partial_n + \alpha)u_2^{k-1}$ on Γ_1 where the parameter α can be optimized to get the fastest convergence. These are a separate class of methods, thus called *optimized Schwarz methods*. More general transmission operators are possible and needed (see, e.g. [94, 98, 121]). The boundary conditions (w.r.t. the subdomain problems) can also be mixed in a Dirichlet-Neumann fashion. For this thesis, we mainly focus on the classical overlapping case.

2.3.2 The Discrete Case

The discrete case is especially relevant since we are mainly interested in using the DD method as a linear solver or as a preconditioner within a Krylov method. Consider the discretized equation $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ in the usual notation with the symmetric and positive definite $\boldsymbol{A} \in \mathbb{R}^n$, the right-hand side $\boldsymbol{b} \in \mathbb{R}^n$, and the solution $\boldsymbol{x} \in \mathbb{R}^n$. In the discrete case, the degrees of freedom (DOFs) act like the volume in the continuous case. So, grouping the DOFs into subdomains is a natural choice, corresponding to a physical domain decomposition in the continuous picture. Following the notation from [98], we denote with $\mathcal{N}$, $|\mathcal{N}| = n$, the set of degrees of freedom, i.e., usually the values on the nodes of a finite element mesh, for example. We can then split these DOFs into $N \ll n$ subsets $\{\mathcal{N}_i\}_{i=1}^N$ such that $\cup_{i=1}^N \mathcal{N}_i = \mathcal{N}$. Another strategy is to first split all DOFs into disjunct sets such that $\mathcal{N}_i' \cap \mathcal{N}_j' = \emptyset$, for $i \neq j$, and then to use the connectivity graph later to enlarge these non-overlapping domains layer by layer iteratively to create the overlapping decompositions $\{\mathcal{N}_i\}_{i=1}^N$. This strategy is a common practice, and software like, e.g., METIS [161] or Scotch [75] can be used. As sketched in Fig. 2.4, such a disjunct decomposition labels the DOFs within the linear system and assigns them to a subdomain.

With a decomposition at hand, we want to formulate an iteration procedure to mimic the parallel Schwarz method (2.74). The solution within one subdomain, i.e., solving a subset of equations within the system, can then be formulated using restriction matrices $\boldsymbol{R}_i \in \{0,1\}^{|\mathcal{N}_i| \times |\mathcal{N}|}$, such that $\boldsymbol{R}_i\boldsymbol{x}$ restricts $\boldsymbol{x}$ to the subdomain $\mathcal{N}_i$ (i.e., Ω_i). Extending a small local vector back to the global vector is then done

using the transpose $\boldsymbol{R}_i^T \in \{0,1\}^{|\mathcal{N}|\times|\mathcal{N}_i|}$, which is called the *extension* matrix. Since we want to operate on the global vector $\boldsymbol{x}$, whose entries (DOFs) belong to multiple subdomains, we need to define also the diagonal *partition of unity* (PU) matrices $\boldsymbol{D}_i \in \mathbb{R}^{|\mathcal{N}_i|\times|\mathcal{N}_i|}$, such that $\boldsymbol{x} = \sum_{i=1}^N \boldsymbol{R}_i^T \boldsymbol{D}_i \boldsymbol{R}_i \boldsymbol{x}$ for all $\boldsymbol{x}$ holds. In the following, we illustrate how these matrices could be constructed.

Example 2.1. Let $\mathcal{N} = \{1,2,3\}$ be split into two subdomains with $\mathcal{N}_1 = \{1,2\}$ and $\mathcal{N}_2 = \{2,3\}$. Then, the restriction matrices and a possible choice of PU matrices are

$$\boldsymbol{R}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad \boldsymbol{R}_2 = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad \boldsymbol{D}_1 = \begin{pmatrix} 1 & 0 \\ 0 & 1/2 \end{pmatrix}, \quad \boldsymbol{D}_2 = \begin{pmatrix} 1/2 & 0 \\ 0 & 1 \end{pmatrix}. \tag{2.75}$$

Let $\boldsymbol{x} = (x_1, x_2, x_3)^T \in \mathbb{R}^3$, then the restriction to the first subdomain is given by $\boldsymbol{R}_1\boldsymbol{x} = (x_1, x_2)^T$. The extension back to the global vector is given by $\boldsymbol{R}_1^T\boldsymbol{R}_1\boldsymbol{x} = (x_1, x_2, 0)^T$. The PU matrices are used to apply a weighting to the DOFs in the overlapping region, i.e., $\boldsymbol{R}_1^T\boldsymbol{D}_1\boldsymbol{R}_1\boldsymbol{x} = (x_1, x_2/2, 0)^T$. The global vector can then be recovered by summing up the contributions from the subdomains, i.e., $\boldsymbol{x} = \boldsymbol{R}_1^T\boldsymbol{D}_1\boldsymbol{R}_1\boldsymbol{x} + \boldsymbol{R}_2^T\boldsymbol{D}_2\boldsymbol{R}_2\boldsymbol{x}$.

After writing out the Jacobi-style iteration (2.74) with the additional weighting from the partition of unity matrices [98], we obtain the fixed-point iteration

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \left(\sum_{i=1}^N \boldsymbol{R}_i^T \boldsymbol{D}_i \left(\boldsymbol{R}_i \boldsymbol{A} \boldsymbol{R}_i^T\right)^{-1} \boldsymbol{R}_i\right) (\boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k). \tag{2.76}$$

Defining the subdomain matrices $\boldsymbol{A}_i := \boldsymbol{R}_i \boldsymbol{A} \boldsymbol{R}_i^T$, the residual in the k-th step $\boldsymbol{r}_k := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k$, and collecting the sum in

$$\boldsymbol{M}_{\mathrm{RAS},1}^{-1} := \sum_{i=1}^N \boldsymbol{R}_i^T \boldsymbol{D}_i \boldsymbol{A}_i^{-1} \boldsymbol{R}_i, \tag{2.77}$$

we can write the iteration in a more compact form as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{M}_{\mathrm{RAS},1}^{-1} \boldsymbol{r}_k. \tag{2.78}$$

This iteration is called the *restricted additive Schwarz method* (RAS) [98] and resembles a preconditioned fixed-point iteration with the one-level preconditioner $\boldsymbol{M}_{\mathrm{RAS},1}^{-1}$. There is also a serial version of the original iteration (2.73) called the *multiplicative Schwarz method* [89] that reads for the simple case of two subdomains case, as

$$\boldsymbol{x}_{k+1/2} = \boldsymbol{x}_k + \boldsymbol{R}_1^T \boldsymbol{A}_1^{-1} \boldsymbol{R}_1 \boldsymbol{r}_k, \tag{2.79}$$

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_{k+1/2} + \boldsymbol{R}_2^T \boldsymbol{A}_2^{-1} \boldsymbol{R}_2 \boldsymbol{r}_{k+1/2}, \tag{2.80}$$

and generalizes to the case of $N > 2$. Elimination of the equations to write them in terms of the global vector $\boldsymbol{x}$ yields a more complicated form (see [89]). Due to the serial application, however, there is no need to use the partition of unity matrices.

Dropping the cross-coupling in the resulting equations to allow for parallelism yields the *additive Schwarz preconditioner* as

$$\boldsymbol{M}_{\mathrm{AS},1}^{-1} = \sum_{i=1}^{N} \boldsymbol{R}_i^T \boldsymbol{A}_i^{-1} \boldsymbol{R}_i, \tag{2.81}$$

for which the fixed-point iteration only converges for a minimal overlap of the subdomains [98]. However, this preconditioner is, in contrast to $\boldsymbol{M}_{\mathrm{RAS},1}$, symmetric and can, thus, be used in combination with the CG method, while the unsymmetric RAS preconditioner needs the GMRES method to be applied [228]. These methods are rarely used in practice as stationary iterations and are almost always used as preconditioners in Krylov solvers as *Krylov acceleration*. The corresponding preconditioned Krylov solvers always perform better or equal to the corresponding fixed-point iteration [89, Thm. 4.1]. In practice, the sums in the preconditioners can be distributed across computers, such that the big problem is split into a set of N problems of reduced size that are, per iteration step, independent. However, for the usual case of a fixed domain Ω and the Poisson problem $-\Delta u = f$ with zero boundary conditions, increasing the number of subdomains N leads to a convergence deterioration [98]. This problem can be overcome with the help of coarse spaces.

2.3.3 Two-Level Methods and Coarse Spaces

Adding a coarse space, i.e., performing a *coarse correction*, adds another level to the domain decomposition method. This correction is done by modifying the preconditioner so that for the RAS2 method, we obtain

$$\boldsymbol{M}_{\mathrm{RAS},2}^{-1} = \boldsymbol{R}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{R}_0 + \boldsymbol{M}_{\mathrm{RAS},1}^{-1}. \tag{2.82}$$

The analog can also be done for the AS preconditioner. In Eq. (2.82), the matrix $\boldsymbol{R}_0^T \in \mathbb{R}^{n \times n_0}$ contains as columns the basis vectors of the n_0-dimensional coarse space, and the coarse space matrix is given by $\boldsymbol{A}_0 = \boldsymbol{R}_0 \boldsymbol{A} \boldsymbol{R}_0^T$. Inverting $\boldsymbol{A}_0$ is typically cheap since $n_0 \ll N \ll n$. There is also a deflated coarse correction for direct use within the stationary iteration (see [3]).

For the Poisson problem, the typical coarse space is based on constant functions. This *Nicolaides* [211] coarse space has the basis vectors $\boldsymbol{v}_i = \boldsymbol{R}_i^T \boldsymbol{D}_i \boldsymbol{R}_i \mathbf{1}$ ($\mathbf{1}$ denoting the vector of ones), for the subdomains $i = 1, \ldots, N$, which are used to construct $\boldsymbol{R}_0^T = (\boldsymbol{v}_1, \ldots, \boldsymbol{v}_N)$ [98]. This coarse space takes care of the low-energy eigenfunctions. For the heterogeneous diffusion problem of the form $-\nabla \cdot (A(x) \nabla u(x)) = f(x)$, this coarse space is not suitable anymore. One can use the *GenEO* [235] coarse space that uses generalized eigenfunctions "in the overlap". Following the original paper [235] for the finite element case, we denote the local "Neumann" matrices with $\boldsymbol{A}_i^{\mathrm{N}}$, obtained only by assembling over elements within the domain Ω_i. Then, the eigenvectors $\boldsymbol{p}_i^{(k)}$ are computed as solutions to the generalized eigenvalue problem as

$$\boldsymbol{A}_i^{\mathrm{N}} \boldsymbol{p}_i^{(k)} = \lambda_i^{(k)} \boldsymbol{D}_i \boldsymbol{A}_i^{\circ} \boldsymbol{D}_i \boldsymbol{p}_i^{(k)}, \tag{2.83}$$

where $\boldsymbol{A}_i^\circ$ corresponds to the local matrix obtained by assembling the elements in the overlap. Note that $\boldsymbol{A}_i^{\mathrm{N}}$ can also replace $\boldsymbol{A}_i^\circ$ [36] in a slightly different formulation. For a threshold $\tau > 0$, the coarse space then collects from each domain Ω_i all eigenvectors with corresponding eigenvalues $\lambda_i^{(k)} \leq \tau^{-1}$, and then uses all $\boldsymbol{R}_i^T \boldsymbol{D}_i \boldsymbol{p}_i^{(k)}$ as a basis of the coarse space, i.e., as columns of $\boldsymbol{R}_0^T$. Such a spectral coarse space approach was also applied in different contexts [237] and extended to the fully algebraic case [2, 3, 130]. A multi-level version was proposed in [36], also presenting alternative formulations for the local eigenvalue problems. An abstract theoretical framework is also given in the original article [235]. If enough eigenfunctions are included, one gets a condition number bound of the form $\kappa(\boldsymbol{M}_{\mathrm{AS},2}^{-1}\boldsymbol{A}) \leq C(1 + H/\delta)$ where H depends on the subdomain size and δ measures the overlap thickness (see [235] for a more precise statement). We discuss the abstract framework in Chapter 4 when analyzing a unique coarse space needed for the Schrödinger eigenvalue problem.

We note that the GenEO space is not the only possible way to construct a coarse space. There are approaches based on Dirichlet-to-Neumann (DtN) maps [206], harmonic enrichment [123], other multiscale approaches [140, 141], and many more. We note here that a very close connection exists to the approaches in the context of multiscale finite element methods.

QOSI: A Quasi-Optimal Factorization Preconditioner for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains

3

This chapter provides a provably quasi-optimal preconditioning strategy of the linear Schrödinger eigenvalue problem with periodic potentials for a possibly non-uniform spatial expansion of the domain. The quasi-optimality is achieved by having the iterative eigenvalue algorithms converge in a constant number of iterations for different domain sizes. In the analysis, we derive an analytic factorization of the spectrum and asymptotically describe it using concepts from the homogenization theory. This decomposition allows us to express the eigenpair as an easy-to-calculate cell problem solution combined with an asymptotically vanishing remainder. We then prove that the easy-to-calculate limit eigenvalue can be used in a shift-and-invert preconditioning strategy to bound the number of eigensolver iterations uniformly. Several numerical examples illustrate the effectiveness of this quasi-optimal preconditioning strategy.

—

This chapter has been published in the article [239]:

- B. Stamm and L. Theisen. "A Quasi-Optimal Factorization Preconditioner for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains". In: *SIAM J. Numer. Anal.* 60.5 (2022), pp. 2508–2537. DOI: 10.1137/21M1456005

3.1 Introduction

This chapter considers the spectral problem for linear time-independent Schrödinger-type operators. Let us consider a parametrized family of d-dimensional boxes Ω_L given by

$$\boldsymbol{z} \in \Omega_L = (0, L)^p \times (0, \ell)^q = \Omega_{\boldsymbol{x}} \times \Omega_{\boldsymbol{y}} \subset \mathbb{R}^d, \tag{3.1}$$

with coordinates $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y}) = (x_1, \dots, x_p, y_1, \dots, y_q)$ and dimensions $L, \ell \in \mathbb{R}$. Note that we provide an extension to arbitrary domains in Section 3.4.3 and only keep the box shape for the early sections of the analysis. We denote by $H_0^1(\Omega_L)$ the standard Sobolev space of index 1 with zero Dirichlet trace on Ω_L.

Then we consider the eigenvalue problem: Find $(\phi_L, \lambda_L) \in (H_0^1(\Omega_L) \setminus \{0\}) \times \mathbb{R}$, such that

$$-\Delta\phi_L + V\phi_L = \lambda_L\phi_L \quad \text{in } \Omega_L. \tag{3.2}$$

Here, ϕ_L and λ_L are the eigenfunctions and -values, respectively, while the function V encodes an external potential applied to the system. Typically, we are interested in computing some of the smallest eigenpairs of Eq. (3.2).

For the analysis, we make the following assumptions about the potential:

(A1) The potential V is directional-periodic with a period of 1 in each expanding direction: $V(\boldsymbol{x}, \boldsymbol{y}) = V(\boldsymbol{x} + \boldsymbol{i}, \boldsymbol{y}) \quad \forall (\boldsymbol{x}, \boldsymbol{y}) \in \Omega_L, \boldsymbol{i} \in \mathbb{Z}^p$;

(A2) The potential V is essentially bounded: $V \in L^\infty(\Omega_L)$;

(A3) The potential V is non-negative: $V \geq 0$ a.e. in Ω_L.

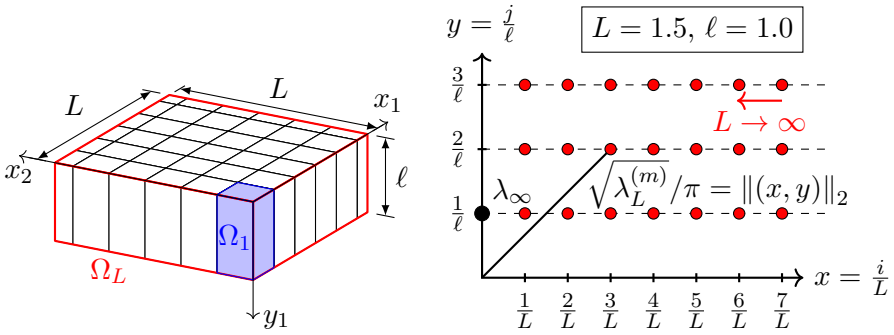
Note that under the assumption (A2), we can always apply a constant spectral shift to the potential without affecting the eigenfunctions to fulfill (A3). Of course, (A1) is only chosen for simplicity and arbitrary periods are possible. Furthermore, we could extend the theory to general elliptic operators satisfying the properties of Section 3.2.1. Fig. 3.1a presents the geometrical framework.

We focus on the case of q fixed dimensions of length ℓ. In contrast, the other p dimensions expand with $L \to \infty$. This geometric setup allows us to study chain-like $(d = 2, p = 1)$ or plane-like $(d = 3, p = 2)$ domains with $L \to \infty$. These are the most common application cases. However, the setup is not limited to $d \leq 3$, and all results also hold in the general case.

Suppose one aims to solve for the ground-state eigenpair (smallest eigenvalue). In that case, the convergence rate of typical numerical algorithms [31, p53] depends on the fundamental ratio between the first and the second $(\lambda_L^{(1)} < \lambda_L^{(2)})$ eigenvalue

$$r_L := |\lambda_L^{(1)}|/|\lambda_L^{(2)}| < 1. \tag{3.3}$$

Our geometrical setup of Ω_L, with $(0, L)^p \to (0, \infty)^p$ and $(0, \ell)^q$ being fixed, can lead to a collapsing fundamental gap $\lambda_L^{(1)} - \lambda_L^{(2)} \to 0$ and thus $r_L \to 1$ as $L \to \infty$. This will deteriorate the convergence rate in the limit. Therefore, the eigensolver routine needs more and more iterations to converge to a fixed tolerance as L increases. To overcome this problem, we theoretically study the operator's spectrum in Eq. (3.2) to construct a suitable shift-and-invert preconditioner [227, p193], such that the preconditioned fundamental gap $r_L(\sigma) := |\lambda_L^{(1)} - \sigma|/|\lambda_L^{(2)} - \sigma|$ is uniformly bounded above by a constant $C < 1$, for all $L > L^*$. For this strategy to work, we need to choose a shift σ based on the asymptotic behavior of the problem. As it turns out later, the quasi-optimal shift σ has to be the asymptotic eigenvalue $\lambda_\infty := \lim_{L\to\infty} \lambda_L^{(1)}$. Throughout this chapter, we understand quasi-optimality in terms of eigensolver iterations, which belong to $\mathcal{O}(1)$ for all L. This complexity is optimal except for an L-independent multiplicative constant. Also, note that we follow [227, p193] and understand preconditioning in the eigenvalue context as a mechanism to speed up the convergence of an iterative solver by applying a spectral transformation [31, p43].



(a) Geometric setup of Ω_L. (b) Eigenvalue lattice for $\Omega_L \subset \mathbb{R}^2, V = 0$.

Figure 3.1: Fig. 3.1a: Geometric setup with $p = 2$ expanding directions with length $L = 5.5$ and $q = 1$ fixed dimensions with length $\ell = 2$. Fig. 3.1b: The Dirichlet Laplacian spectrum on a rectangle domain $\Omega_L = (0, L) \times (0, l)$ mapped to an eigenvalue lattice.

3.1.1 Motivation: Collapsing Fundamental Gap for the Laplace EVP

Even for the simplest potential satisfying the assumptions (A1)–(A3), namely $V(\boldsymbol{z}) = 0$, the fundamental gap r_L decreases if only a subset of directions in Ω_L is increased. The simplicity of the Laplace eigenvalue problem allows us to highlight the main challenges by considering the explicitly known eigenvalues. For $p = 1$ and $q = 1$, the pure Laplace eigenvalue problem has eigenvalues $\lambda_L^{(1)} = \pi^2/L^2 + \pi^2/\ell^2, \lambda_L^{(2)} = 4\pi^2/L^2 + \pi^2/\ell^2$. It is then evident that $\lim_{L\to\infty} \lambda_L^{(1)}/\lambda_L^{(2)} = 1$ for $L \to \infty$. Thus, this leads to a decreasing convergence rate r_L. The collapsing fundamental gap is visible in the eigenvalue lattice illustrated in Fig. 3.1b. In such a representation, each point represents an eigenvalue $\lambda_L^{(m)}$, and its distance to the origin corresponds to $\mathrm{sqrt}(\lambda_L^{(m)}/\pi)$. All eigenvalues will form continuous x-parallel lines for the asymptotic case of $L \to \infty$. For this Laplace eigenvalue problem, a shift of $\sigma = \lambda_\infty = \pi^2/l^2$ would lead to $\lim_{L\to\infty} r_L(\sigma) = 1/4 < 1$.

This simple example serves as a motivation. However, to give a systematic approach to choosing the asymptotic correct shift σ in the general case, we develop a framework for characterizing the asymptotic behavior of the spectral properties for Schrödinger operators with periodic potentials satisfying (A1)–(A3). Knowing the asymptotic behavior will allow solving the algebraic eigenvalue problem within only a constant number of eigensolver iterations.

3.1.2 State-of-the-Art and Context

We can embed our results into existing research for three different aspects – the considered model equation, the geometrical setup with the present periodic potential, and other mathematical analyses for related equations.

First, the Schrödinger equation Eq. (3.2) describes the stationary states of the wave function ϕ_L for a quantum-mechanical system influenced by an external potential V. Therefore, example applications naturally arise in computational chemistry and quantum mechanics. Since the present model is one of the simpler models, it is only suitable for the direct simulation of basic quantum systems. In more elaborate electronic structure calculations, a nonlinear version of the Schrödinger equation is used. However, there is always a need to solve systems similar to Eq. (3.2) in self-consistent field (SCF) iterations [64, 139, 148] or other iteration schemes to solve these nonlinear Schrödinger-type equations [21, 22, 23]. One of such examples is the Bose–Einstein condensate, either modeled with random or disorder potentials [14, 17] or modeled with the Gross–Pitaevskii eigenvalue problem [16, 17, 18, 146]. Other applications for the same equation Eq. (3.2) arise in studying the power distribution in a nuclear reactor core [8, 10, 114].

Second, when it comes to the geometric setup of only a subset of dimensions expanding, applications arise, for example, in material science to study the electronic properties of plane-like, layered [69, 70] or chain-like structures, such as carbon nanotubes [1] or polymers chains [261].

Third, from a mathematical point of view, the study of elliptic operators in the context of source [76, 77, 80, 83, 100, 101, 241] or eigenvalue problems [61, 270] with homogenization is closely related since V is periodic (at least directional in our setup). Also, for eigenvalue problems with periodic coefficients, results in [79, 81, 82] show the presence of an asymptotic limit when the domain expands in some directions to infinity. Finally, from a technical point of view, our analysis in Section 3.2 extends aspects of the work by Allaire et al. in [6, 7, 8, 9, 10, 11]. Especially the concept of factorization will be one of the main techniques in our analysis. It allows us to analytically describe the spectrum of the system in terms of easier cell problems. This idea traces back to [164, 165, 257].

3.1.3 Contribution and Main Results

This chapter aims to provide a numerical framework to solve the eigenvalue problem Eq. (3.2) in a fixed number of eigensolver iterations for all domain sizes $L \to \infty$. We, therefore, propose a shift-and-invert strategy with a quasi-optimally chosen shift. The theoretical derivation of this particular shift is based on the following factorization of the eigenfunctions (see Theorem 3.1 for a more precise statement)

$$\phi_L^{(m)} = \psi \cdot u_{y,1} \cdot u_{y,2}^{(m)} \qquad = \varphi_y \cdot u_{y,2}^{(m)}, \tag{3.4}$$

$$\lambda_L^{(m)} = \lambda_\psi + \lambda_{u_{y,1}} + \lambda_{u_{y,2}}^{(m)} = \lambda_{\varphi_y} + \lambda_{u_{y,2}}^{(m)} = \lambda_{\varphi_y} + \mathcal{O}(1/L^2). \tag{3.5}$$

The above characterization highlights that we can simply use λ_{φ_y} as the quasi-optimal shift since we will show that the remaining term $\lambda_{u_{y,2}}^{(m)}$ tends to zero as $L \to \infty$ for all m. The $\mathcal{O}(1/L^2)$ contribution in Eq. (3.5) is not uniform in m since it depends on the m-th eigenvalue of a homogenized equation, as shown later in Theorem 3.2. In contrast to existing literature, this statement considers the case where only a subset

of dimensions expands, and the periodicity is directional, which is essential given the potential practical applications. The eigenpair $(\varphi_y, \lambda_{\varphi_y})$ can be obtained in a constant time since it does not depend on $L \to \infty$ as it is a solution to a fixed-size spectral cell problem. We then show that this eigenpair is the asymptotic limit as

$$\lim_{L\to\infty} \lambda_L^{(m)} = \lambda_{\varphi_y}. \tag{3.6}$$

These results, then, directly imply that the preconditioned fundamental ratio r_L is uniformly bounded from above by a constant for all $L > L^*$, which is smaller than one (see Theorem 3.3 for a more precise statement). Since the convergence speed of the iterative eigensolvers depends on precisely this ratio, they converge in a constant number of iterations, and our goal is achieved.

The main challenges arise in the analysis and the quasi-optimality proof of the preconditioner. Using factorization results to construct a quasi-optimal preconditioner for an anisotropic domain size increase, which can be computed in $\mathcal{O}(1)$, is not covered, up to our knowledge, in the existing literature. Moreover, although the idea of factorization is not new (see the references in Section 3.1.2), it was not yet applied in the context of weighted and thus potentially degenerate Sobolev spaces. However, precisely this setup of degenerate weights is necessary for our quasi-optimality analysis since the zero Dirichlet boundary conditions on the $\boldsymbol{y}$-boundary significantly contribute to the asymptotic behavior of the spectrum. In addition, proving quasi-optimality also requires uniform bounds of the preconditioned system's fundamental ratio. Thus, another challenge is interpreting the expanding problem as a homogenization problem in a degenerate situation. This observation allows us to explicitly determine the asymptotic behavior of the m-dependent contribution in Eqs. (3.4) and (3.5). However, unlike the classical homogenization theory, our homogenized limit is purely determined by an p- rather than an $(p+q)$-dimensional problem, which seems to be a unique specialty of anisotropic expansion problems.

3.1.4 Outline of the Chapter

In Section 3.2, we present the theoretical framework for the factorization approach in Section 3.2.2, which allows us to consider the remaining simplified problems using the theory of homogenization in Section 3.2.3 to derive quasi-optimality statements. We then discretize and solve the eigensystem in Section 3.3 and show that the theoretical results also hold when specific subspace properties are met in the discrete setting. Next, Section 3.4 presents various numerical examples and shows the relevance of the method for solving practical problems. Finally, we conclude with some remarks and point out future work in Section 3.5.

3.2 Factorization and Homogenization of the Model Problem

This section will use factorization and homogenization to derive the asymptotic spectrum, which allows us to specify the limit eigenvalue $\lambda_{\varphi_y} = \lim_{L\to\infty} \lambda_L^{(m)}$.

3.2.1 Existence and Regularity Results

The second-order partial differential operator in Eq. (3.2) is self-adjoint (by the symmetry of the diffusion matrix δ_{ij}), is positive-definite (by ellipticity/coercivity of the Laplacian plus a non-negative potential), and has bounded coefficients (since $V \in L^\infty(\Omega_L)$). We can therefore recall the classical existence results from [127, 146, 257] to establish the well-posedness of our problem, that there exists a sequence ($m = 1, \dots, \infty$) of eigenvalues with finite multiplicity $\lambda_L^{(m)}$ and a sequence of eigenfunctions $\phi_L^{(m)}$ (orthogonal basis of $H_0^1(\Omega_L)$) such that $0 < \lambda_L^{(1)} < \lambda_L^{(2)} \leq \lambda_L^{(3)} \leq \dots \to \infty$ and $\phi_L^{(1)} > 0$ a.e. in Ω_L.

3.2.2 Factorization of the Eigenfunctions and Eigenvalues

Our next step is to establish factorizations to solve the eigenvalue problem Eq. (3.2). Thus, we describe splitting an eigenfunction into a product of two or more functions and splitting an eigenvalue into the sum of two or more eigenvalues. This splitting can be seen as a generalization to the separation of variables for the pure Laplacian case.

Let $\mathcal{D}(\Omega_L) = C_c^\infty(\Omega_L)$ be the space of compactly supported test functions on Ω_L and ρ a weight function (measurable and positive a.e. in Ω_L). We then use the weighted Sobolev [178, 257] spaces as

$$H^1(\Omega_L;\rho) = \left\{ u \in \mathcal{D}'(\Omega_L) \;\middle|\; \|u\|_{H^1(\Omega_L;\rho)} < \infty \right\}, \tag{3.7}$$

$$H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho) = \left\{ u \in H^1(\Omega_L;\rho) \;\middle|\; \begin{cases} \mathcal{B}_{\boldsymbol{x}}(u) = 0 \text{ on } \partial(0,L)^p \times (0,\ell)^q \\ \mathcal{B}_{\boldsymbol{y}}(u) = 0 \text{ on } (0,L)^p \times \partial(0,\ell)^q \end{cases} \right\}, \tag{3.8}$$

for some general boundary operators $\mathcal{B}_{\boldsymbol{x}}, \mathcal{B}_{\boldsymbol{y}}$, which are equipped with the weighted norm

$$\|\cdot\|_{H^1(\Omega_L;\rho)} = \sqrt{\|\nabla \cdot\|^2_{L^2(\Omega_L;\rho)} + \|\cdot\|^2_{L^2(\Omega_L;\rho)}}, \tag{3.9}$$

using $\|\cdot\|^2_{L^2(\Omega_L;\rho)} := \|\sqrt{\rho}\cdot\|^2_{L^2(\Omega_L)}$ in the classical L^2-sense. The corresponding scalar product is $\langle\cdot,\cdot\rangle_{L^2(\Omega_L;\rho)} = \langle\rho\cdot,\cdot\rangle_{L^2(\Omega_L)}$. For the boundary operators, we use $\mathcal{B}_{\boldsymbol{x}}, \mathcal{B}_{\boldsymbol{y}} \in \{\mathcal{B}_d, \mathcal{B}_n, \mathcal{B}_\#\}$ with

$$\text{Dirichlet: } \mathcal{B}_d(u)(\boldsymbol{z}) = u(\boldsymbol{z}), \tag{3.10}$$

$$\text{Neumann: } \mathcal{B}_n(u)(\boldsymbol{z}) = \rho(\boldsymbol{z})\nabla u(\boldsymbol{z}) \cdot \boldsymbol{n}(\boldsymbol{z}), \tag{3.11}$$

$$\text{Periodic: } \mathcal{B}_\#(u)(\boldsymbol{z}) = u(\boldsymbol{z}) - u\left((z_i - n_i(\boldsymbol{z})L)_{i=1}^p, (z_i - n_i(\boldsymbol{z})\ell)_{i=p+1}^q\right), \tag{3.12}$$

where the unit normal-vector is denoted by $\boldsymbol{n}(\boldsymbol{z})$ for $\boldsymbol{z} \in \partial\Omega_L$.

In the following, we use multiple eigenvalue problems and their solutions. Therefore, we unify the notation and introduce the abstract notation:

Definition 3.1 (Prototype of a Schrödinger eigenvalue problem)**.** For $\Omega_L = (0,L)^p \times (0,\ell)^q \subset \mathbb{R}^d$, $0 \leq \rho, V \in L^\infty(\Omega_L)$, and $1/\rho \in L^1_{\text{loc}}(\Omega_L)$, we define

$$\left(u^{(m)}_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}},\rho,V}(\Omega_L), \lambda^{(m)}_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}},\rho,V}(\Omega_L)\right), \tag{3.13}$$

to be the m-th eigenpair (including multiplicities) of the generalized Schrödinger-type eigenvalue problem: Find $(u, \lambda) \in (H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho) \setminus \{0\}) \times \mathbb{R}$, such that

$$-\nabla \cdot (\rho \nabla u) + Vu = \lambda \rho u \text{ in } \Omega_L. \tag{3.14}$$

Remark 3.1. If the weight ρ is zero only at the boundary of Ω_L, then we have $1/\rho \in L^1_{\text{loc}}(\Omega_L)$, which implies that $H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho)$ is a Banach space [178, p235]. On the other hand, in the case of ρ being bounded from above and uniformly positive, i.e., $0 < \mathsf{c} < \rho < \mathsf{C}$ a.e. in Ω_L, the ρ-weighted Sobolev space from Eq. (3.7) is equivalent to the classical Sobolev space $H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho) = H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L)$, and we omit ρ in the notation.

Remark 3.2 (Weak form). The corresponding weak formulation of Eq. (3.14) reads: Find $(u, \lambda) \in (H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho) \setminus \{0\}) \times \mathbb{R}$ such that

$$\forall v \in H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho): \quad \int_{\Omega_L} \rho \nabla u \cdot \nabla v \, \mathrm{d}\boldsymbol{z} + \int_{\Omega_L} Vuv \, \mathrm{d}\boldsymbol{z} = \lambda \int_{\Omega_L} \rho uv \, \mathrm{d}\boldsymbol{z}. \tag{3.15}$$

Remark 3.3 (Min-max characterization). Since the weight ρ is a scalar function, Eq. (3.14) is self-adjoint for the presented boundary conditions, and we can express the eigenpair through the min-max characterization (c.f. [93, 164]):

$$\lambda^{(m)} = \min_{\substack{W_m \subset H^1_{\mathcal{B}_{\boldsymbol{x}},\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho) \\ \dim W_m = m}} \max_{\substack{u \in W_m \\ u \neq 0}} \mathcal{R}_{\rho,V}(u), \tag{3.16}$$

with the Rayleigh quotient, defined by

$$\mathcal{R}_{\rho,V}(u) = \frac{\int_{\Omega_L} \rho(\boldsymbol{z}) \nabla u(\boldsymbol{z}) \cdot \nabla u(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} + \int_{\Omega_L} V(\boldsymbol{z}) u^2(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}}{\int_{\Omega_L} \rho(\boldsymbol{z}) u^2(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z}}, \tag{3.17}$$

which is identical for all the considered boundary conditions $\mathcal{B}_{\boldsymbol{x}}, \mathcal{B}_{\boldsymbol{y}} \in \{\mathcal{B}_d, \mathcal{B}_n, \mathcal{B}_\#\}$ of Definition 3.1.

We define $E^{\#}_{\boldsymbol{x}} : H^1_{\mathcal{B}_\#,\mathcal{B}_{\boldsymbol{y}}}(\Omega_1;\rho) \to H^1_{\mathcal{B}_\#,\mathcal{B}_{\boldsymbol{y}}}(\Omega_L;\rho)$ as the periodic extension operator in the $\boldsymbol{x}$-direction for an $\boldsymbol{x}$-periodic weight ρ and $\mathcal{B}_{\boldsymbol{y}} \in \{\mathcal{B}_d, \mathcal{B}_n, \mathcal{B}_\#\}$. We are now prepared to state our first main theoretical result:

Theorem 3.1 (Factorzation of eigenfunctions and summation of eigenvalues)**.** *The m-th eigenfunction of the Schrödinger eigenvalue problem Eq. (3.2) can be factorized into*

$$u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,1,V} = \psi \cdot u^{(m)} = \psi \cdot u_{x,1} \cdot u^{(m)}_{x,2} = \varphi_x \cdot u^{(m)}_{x,2} \tag{3.18}$$

$$= \psi \cdot u_{y,1} \cdot u^{(m)}_{y,2} = \varphi_y \cdot u^{(m)}_{y,2} \tag{3.19}$$

while the m-th eigenvalue can be summed correspondingly as

$$\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,1,V} = \lambda_\psi + \lambda^{(m)}_u = \lambda_\psi + \lambda_{u_{x,1}} + \lambda^{(m)}_{u_{x,2}} = \lambda_{\varphi_x} + \lambda^{(m)}_{u_{x,2}} \tag{3.20}$$

$$= \lambda_\psi + \lambda_{u_{y,1}} + \lambda^{(m)}_{u_{y,2}} = \lambda_{\varphi_y} + \lambda^{(m)}_{u_{y,2}} \tag{3.21}$$

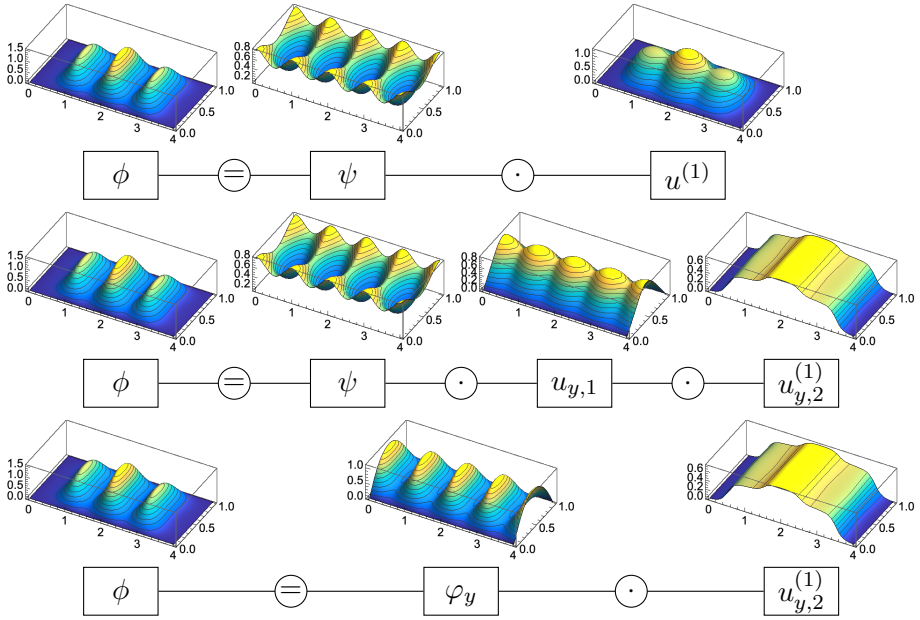


Figure 3.2: Visualization of the factorization for the ground state solution of $-\Delta\phi + V\phi = \lambda\phi, \phi = 0$ on $\partial\Omega_4$ with $V(x,y) = 10^2(\sin x)^2(\sin y)^2$.

where

$$\psi = E_{\boldsymbol{x}}^{\#} u^{(1)}_{\mathcal{B}_{\#},\mathcal{B}_{\#},1,V}(\Omega_1),\ u^{(m)} = u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0}(\Omega_L), \tag{3.22}$$

$$u_{x,1} = u^{(1)}_{\mathcal{B}_d,\mathcal{B}_{\#},\psi^2,0}(\Omega_L), \quad u^{(m)}_{x,2} = u^{(m)}_{\mathcal{B}_n,\mathcal{B}_d,\varphi_x^2,0}(\Omega_L),\ \varphi_x = u^{(1)}_{\mathcal{B}_d,\mathcal{B}_{\#},1,V}(\Omega_L), \tag{3.23}$$

$$u_{y,1} = E_{\boldsymbol{x}}^{\#} u^{(1)}_{\mathcal{B}_{\#},\mathcal{B}_d,\psi^2,0}(\Omega_1),\ u^{(m)}_{y,2} = u^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\varphi_y^2,0}(\Omega_L),\ \varphi_y = E_{\boldsymbol{x}}^{\#} u^{(1)}_{\mathcal{B}_{\#},\mathcal{B}_d,1,V}(\Omega_1). \tag{3.24}$$

A graphical representation of Theorem 3.1 is presented in Fig. 3.2 for $m = 1$, where the scale separation of, e.g., $\varphi_y^{(1)}$ into a short and $u_{y,2}^{(1)}$ into a large scale is visible. Moreover, the first excited eigenfunctions with $m \in \{2,3,4\}$ are visualized in Fig. 3.3. Herein, the m-dependence entirely goes into the $u_{y,2}^{(m)}$ function for the excited eigenfunctions since $\varphi_y^{(1)}$ is fixed.

To prove Theorem 3.1, we need to show that the factorizations are valid changes of variables. The application of the first *factorization principle* in Eq. (3.18), i.e.,

$$u^{(m)} = u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0} = u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}/\psi, \tag{3.25}$$

removes the potential V from the eigenvalue problem while still encoding the corresponding information through ψ^2. The inducing function $\psi = E_{\boldsymbol{x}}^{\#} u^{(1)}_{\mathcal{B}_{\#},\mathcal{B}_{\#},1,V}(\Omega_1)$ is the solution to a spectral cell problem, and it was shown in [8] that this factorization is indeed a diffeomorphism in $H_0^1(\Omega_L)$. Such factorization operators will apply a change of variables in the min-max characterization of the eigenvalue.

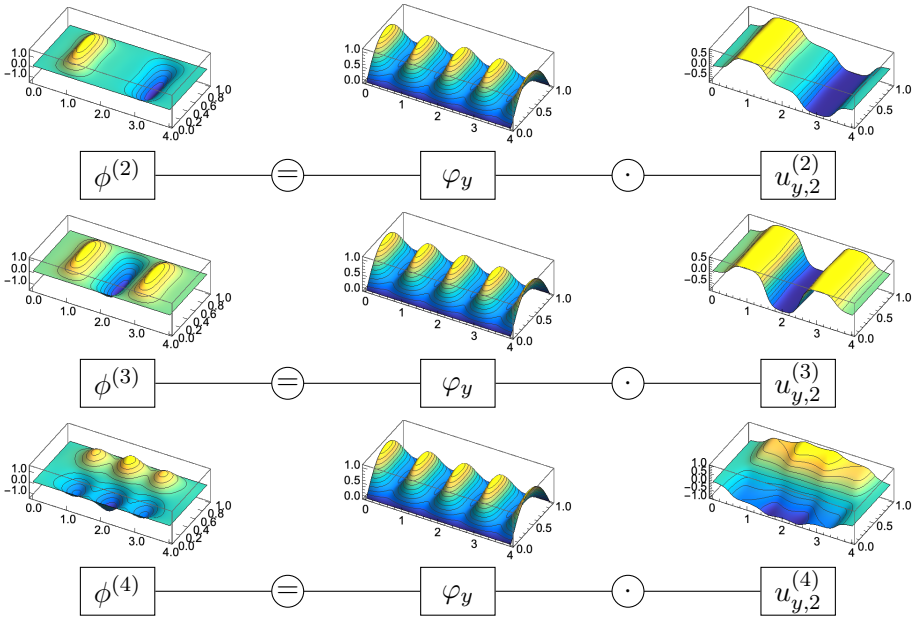


Figure 3.3: Visualization of the factorization for some excited states of $-\Delta\phi^{(m)} + V\phi^{(m)} = \lambda^{(m)}\phi^{(m)}$ with $V(x,y) = 10^2(\sin x)^2(\sin y)^2$: By construction, the m-dependence entirely goes into the $u_{y,2}$ contribution.

In contrast to Eq. (3.25) where ψ^2 is bounded from below a.e. by a positive constant, we will also need factorization operators induced by functions tending to zero at boundary parts due to homogeneous Dirichlet boundary conditions. In such cases, we need to adapt the factorization principle as the boundedness of the division operator is not directly visible, which makes the analysis much more subtle. Thus, we need:

Lemma 3.1 (Factorization operator with degeneracy and singularity)**.** *Let the inducing function* $u_{y,1} := E_{\boldsymbol{x}}^{\#} u^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,\psi^2,0}(\Omega_1) \in H^1_{\mathcal{B}_\#,\mathcal{B}_d}(\Omega_L)$ *with* $0 < \mathsf{c} < \psi^2 < \mathsf{C}$ *a.e. be given. Then, the linear factorization operator defined by*

$$\begin{aligned} T : H^1_{\mathcal{B}_d,\mathcal{B}_d}(\Omega_L) &\to H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u_{y,1}^2) \\ u \mapsto T(u) := \boldsymbol{z} \mapsto u(\boldsymbol{z})/u_{y,1}(\boldsymbol{z}) &\text{ a.e. in } \Omega_L \end{aligned} \tag{3.26}$$

is bi-continuous and thus a diffeomorphism.

Proof. Noting that T is a division operator, the corresponding multiplication operator T^{-1} is the (left and right) inverse operator. For a simpler notation, we use $H^1_0(\Omega_L) = H^1_{\mathcal{B}_d,\mathcal{B}_d}(\Omega_L)$. So we study

$$\begin{aligned} \tilde{T} : H^1_0(\Omega_L) &\to W & \tilde{T}^{-1} : W &\to H^1_0(\Omega_L) \\ u \mapsto u_{y,2} := \tilde{T}(u) &= \frac{u}{u_{y,1}} & u_{y,2} \mapsto u := \tilde{T}^{-1}(u_{y,2}) &= u_{y,1}u_{y,2} \end{aligned} \tag{3.27}$$

with the abstract set $W := \mathrm{Im}(\tilde{T}) = \mathrm{Dom}(\tilde{T}^{-1})$ as

$$W = \left\{ u_{y,2} \in \mathcal{D}'(\Omega_L) \;\middle|\; \exists u \in H^1_0(\Omega_L) : u_{y,2} = \tilde{T}(u) \right\}. \tag{3.28}$$

We show that the image of $\tilde{T}$ is the $u^2_{y,1}$-induced space $H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u^2_{y,1})$. We have $W \subset \mathcal{D}'(\Omega_L)$ since $1/u_{y,1} \in L^1_{\mathrm{loc}}(\Omega_L)$ as $u_{y,1} = 0$ only at the $\boldsymbol{y}$-boundary, and therefore $\forall \phi \in \mathcal{D}(\Omega_L) = C^\infty_c(\Omega_L)$, $\tilde{T}(\phi) = \int_{\Omega_L} \frac{1}{u_{y,1}} \phi < \|\phi\|_\infty \int_{\mathrm{supp}(\phi)} \frac{1}{u_{y,1}} < \infty$ is a linear bounded functional since $\mathrm{supp}(\phi)$ is compact. Then we have $W \subset H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u^2_{y,1})$ since $u_{y,2} \in W$ implies that there exists an $u \in H^1_0(\Omega_L)$, such that $u_{y,2} = \frac{u}{u_{y,1}}$, which yields $\mathcal{B}_d(u_{y,2}) = 0$ on $\partial\Omega_{\boldsymbol{x}}$ and trivially fulfilled $\mathcal{B}_n(u_{y,2}) = 0$ on $\partial\Omega_{\boldsymbol{y}}$, and allows us to take

$$\tilde{T}^{-1}(u_{y,2}) = u_{y,1}u_{y,2} = \frac{u_{y,1}}{u_{y,1}} u = u \in H^1_0(\Omega_L) \Rightarrow u_{y,2} \in H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u^2_{y,1}). \tag{3.29}$$

We also have $H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u^2_{y,1}) \subset W$ since $u_{y,2} \in H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u^2_{y,1})$ implies $\tilde{T}^{-1}(u_{y,2}) = u_{y,1}u_{y,2} = u \in H^1_0(\Omega_L)$ such that

$$\exists u \in H^1_0(\Omega_L) : u_{y,2} = \frac{u}{u_{y,1}} = \tilde{T}(u) \Rightarrow u_{y,2} \in W. \tag{3.30}$$

Thus $W = H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u^2_{y,1})$, so $\tilde{T}$ defined by from Eq. (3.27) coincides with T from Eq. (3.26). Moreover, the weighted Sobolev space $H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L; u^2_{y,1})$ is a Banach space since $u_{y,1}$ only degenerates on the boundary (c.f. Remark 3.1).

Since T and T^{-1} are both surjective, T is bijective. We now show the continuity of T^{-1} as this is the more straightforward direction being a multiplication operator. Since all $u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0}$ form a basis of $H^1_0(\Omega_L)$ (c.f. Section 3.2.1), it suffices to show continuity for all basis functions $u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0}$ and to conclude by the linearity of the operator T^{-1}. Thus, let m be fixed, $u := u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0}$, and $u_{y,2} := T(u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0})$. Then it follows for $u = T^{-1}(u_{y,2}) = u_{y,1}u_{y,2}$ that

$$\int_{\Omega_L} \psi^2 \nabla u \cdot \nabla u \, \mathrm{d}\boldsymbol{z} = \int_{\Omega_L} \psi^2 \left(u^2_{y,1} \nabla u_{y,2} \cdot \nabla u_{y,2} + \nabla u_{y,1} \cdot \nabla (u^2_{y,2} u_{y,1}) \right) \mathrm{d}\boldsymbol{z}, \tag{3.31}$$

which is well defined by the assumption that u is the corresponding eigenfunction for the first expression in Eq. (3.31). We further have for the last term in Eq. (3.31) that

$$\begin{aligned} \int_{\Omega_L} \psi^2 \nabla u_{y,1} \cdot \nabla (u^2_{y,2} u_{y,1}) \, \mathrm{d}\boldsymbol{z} &= -\int_{\Omega_L} \nabla \cdot \left(\psi^2 \nabla u_{y,1} \right) u^2_{y,2} u_{y,1} \, \mathrm{d}\boldsymbol{z} \\ &= \lambda^{(1)}_{u_{y,1}} \int_{\Omega_L} \left(\psi^2 u_{y,1} \right) \left(u^2_{y,2} u_{y,1} \right) \mathrm{d}\boldsymbol{z}, \end{aligned} \tag{3.32}$$

by the definition of $u_{y,1} = E^{\#}_{\boldsymbol{x}} u^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,\psi^2,0}(\Omega_1)$ according to Eq. (3.14) multiplied with $u^2_{y,2}u_{y,1}$. Together with $0 < \mathsf{c} < \psi^2 < \mathsf{C}$, it then follows from Eq. (3.31) that

$$\begin{aligned} \mathsf{c} \int_{\Omega_L} \nabla u \cdot \nabla u \, \mathrm{d}\boldsymbol{z} &\leq \mathsf{C} \int_{\Omega_L} u^2_{y,1} \nabla u_{y,2} \cdot \nabla u_{y,2} \, \mathrm{d}\boldsymbol{z} + \mathsf{C} \lambda^{(1)}_{u_{y,1}} \int_{\Omega_L} u^2_{y,1} u^2_{y,2} \, \mathrm{d}\boldsymbol{z} \\ &\leq \mathsf{C} \max\{1, \lambda^{(1)}_{u_{y,1}}\} ||u_{y,2}||^2_{H^1(\Omega_L; u^2_{y,1})}. \end{aligned} \tag{3.33}$$

Adding $||u||^2_{L^2(\Omega_L)} = ||u_{y,2}||^2_{L^2(\Omega_L;u^2_{y,1})}$ on both sides of Eq. (3.33) yields

$$||\nabla u||^2_{L^2(\Omega_L)} + ||u||^2_{L^2(\Omega_L)} \leq \frac{\mathsf{C}\max\{1,\lambda^{(1)}_{u_{y,1}}\}}{\mathsf{c}}||u_{y,2}||^2_{H^1(\Omega_L;u^2_{y,1})} + ||u_{y,2}||^2_{L^2(\Omega_L;u^2_{y,1})}, \tag{3.34}$$

which finally, with $u_{y,2} = T(u)$, provides us that $||u||^2_{H^1(\Omega_L)} \leq \mathsf{D}||T(u)||^2_{H^1(\Omega_L;u^2_{y,1})}$ for some $\mathsf{D} > 0$. This is equivalent to

$$||T^{-1}(u_{y,2})||^2_{H^1(\Omega_L)} \leq \mathsf{D}||u_{y,2}||^2_{H^1(\Omega_L;u^2_{y,1})}. \tag{3.35}$$

As T^{-1} is a linear, bijective, and continuous (by Eq. (3.35)) operator between two Banach spaces, the inverse $(T^{-1})^{-1} = T$ is also continuous [52, p35] with $||T(u)||^2_{H^1(\Omega_L;u^2_{y,1})} \leq \tilde{\mathsf{D}}||u||^2_{H^1(\Omega_L)}$. Bi-continuity of T implies that T and T^{-1} are continuously Fréchet-differentiable since they are both linear. Thus, T is a diffeomorphism [118]. □

In order to apply the above-defined factorization operator in the min-max setting, we have the following:

Lemma 3.2 (Rayleigh Quotients after Factorization)**.** *Let $\phi^{(m)} = u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L)$ and $u^{(m)}, \psi$ be given as in Theorem 3.1. Then, after the factorization of $\phi^{(m)} = u^{(m)} \cdot \psi$, the corresponding Rayleigh quotient reads*

$$\mathcal{R}_{1,V}(\phi^{(m)}) = \mathcal{R}_{\psi^2,0}\left(u^{(m)}\right) + \lambda_\psi. \tag{3.36}$$

Proof. We first note that the factorization allows for the splitting

$$\begin{aligned}\nabla\phi^{(m)} \cdot \nabla\phi^{(m)} &= \nabla\left(u^{(m)}\psi\right) \cdot \nabla\left(u^{(m)}\psi\right) \\ &= \psi^2\nabla u^{(m)} \cdot \nabla u^{(m)} + \nabla\psi \cdot \nabla\left(\left(u^{(m)}\right)^2\psi\right).\end{aligned} \tag{3.37}$$

Using the splitting Eq. (3.37), we obtain that $\mathcal{R}_{1,V}(\phi^{(m)})$ is equal to

$$\begin{aligned}&\frac{\int_{\Omega_L}\psi^2\nabla u^{(m)} \cdot \nabla u^{(m)}\,\mathrm{d}\boldsymbol{z}}{\int_{\Omega_L}\psi^2\left(u^{(m)}\right)^2\,\mathrm{d}\boldsymbol{z}} + \frac{\int_{\Omega_L}\left(\nabla\psi \cdot \nabla\left(\left(u^{(m)}\right)^2\psi\right) + V\psi\left(\left(u^{(m)}\right)^2\psi\right)\right)\mathrm{d}\boldsymbol{z}}{\int_{\Omega_L}\psi\left(\left(u^{(m)}\right)^2\psi\right)\mathrm{d}\boldsymbol{z}} \\ &= \mathcal{R}_{\psi^2,0}\left(u^{(m)}\right) + \frac{\int_{\Omega_L}\left(-\nabla\cdot(\nabla\psi) + V\psi\right)\left(u^{(m)}\right)^2\psi\,\mathrm{d}\boldsymbol{z}}{\int_{\Omega_L}\psi\left(\left(u^{(m)}\right)^2\psi\right)\mathrm{d}\boldsymbol{z}} \\ &= \mathcal{R}_{\psi^2,0}\left(u^{(m)}\right) + \frac{\int_{\Omega_L}\lambda_\varphi\psi\left(\left(u^{(m)}\right)^2\psi\right)\mathrm{d}\boldsymbol{z}}{\int_{\Omega_L}\psi\left(\left(u^{(m)}\right)^2\psi\right)\mathrm{d}\boldsymbol{z}} = \mathcal{R}_{\psi^2,0}\left(u^{(m)}\right) + \lambda_\psi,\end{aligned} \tag{3.38}$$

by the definition of the eigenfunction $\psi = E^{\#}_{\boldsymbol{x}}u^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,1,V}(\Omega_1)$. □

We can now combine all the above to prove Theorem 3.1:

Proof of Theorem 3.1. For a simpler notation, let $\phi = u^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L)$. We apply the factorization principle twice with

$$\phi = u\psi = (T_\psi)^{-1}(u) \quad \text{and} \quad u = u_{y,1}u_{y,2} = (T_{u_{y,1}})^{-1}(u_{y,2}). \tag{3.39}$$

The operations are defined in Eq. (3.25) and Lemma 3.1, and the latter ensures that these are valid changes of variables in the min-max characterization. With Lemma 3.2 and the first change of variables, we then obtain for $\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L)$ that

$$\begin{aligned}\min_{\substack{W_m \subset H^1_{\mathcal{B}_d,\mathcal{B}_d}(\Omega_L)\\ \dim W_m = m}} \max_{\substack{\phi\in W_m\\ \phi\neq 0}} \mathcal{R}_{1,V}(\phi) &= \min_{\substack{W_m \subset H^1_{\mathcal{B}_d,\mathcal{B}_d}(\Omega_L)\\ \dim W_m = m}} \max_{\substack{u\in W_m\\ u\neq 0}} \mathcal{R}_{\psi^2,0}(u) + \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_\#,1,V}(\Omega_1)\\ &= \lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0}(\Omega_L) + \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_\#,1,V}(\Omega_1).\end{aligned} \tag{3.40}$$

The potential V is now encoded in the diffusion coefficient ψ^2. We now continue in the same fashion with the second factorization and obtain

$$\begin{aligned}\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,\psi^2,0} &= \min_{\substack{W_m \subset H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_L;u^2_{y,1})\\ \dim W_m = m}} \max_{\substack{u_{y,2}\in W_m\\ u_{y,2}\neq 0}} \mathcal{R}_{\psi^2 u^2_{y,1},0}(u_{y,2}) + \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,\psi^2,0}(\Omega_1)\\ &= \lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\psi^2 u^2_{y,1},0}(\Omega_L) + \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,\psi^2,0}(\Omega_1).\end{aligned} \tag{3.41}$$

Here, we used $(T_{u_{y,1}})^{-1}(u_{y,2})$ being a diffeomorphism by Lemma 3.1 and, therefore, a well-defined change of variables. The corresponding eigenfunction multiplication in Eq. (3.19) is a direct result of the factorizations of Eq. (3.39). If we then also apply the ψ-factorization to φ_y, we obtain $\varphi_y = \psi u_{y,1}$ and $\lambda_{\varphi_y} = \lambda_\psi + \lambda_{u_{y,1}}$, which concludes the proof of Eq. (3.21). The other relations, i.e. Eqs. (3.18) and (3.20), follow analogously with applying their respective factorizations $(T(\cdot))^{-1}$ in Eq. (3.39). □

3.2.3 Homogenization in the Expanding Directions

In order to entirely characterize the asymptotic behavior of the spectrum as $L \to \infty$, we will now consider the contribution $\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\psi^2 u^2_{y,1},0}(\Omega_L)$ in Eq. (3.21) as the only one that depends on m after the factorization of Theorem 3.1. Then, we can make a precise statement about the asymptotic behavior of this remainder.

Theorem 3.2 (Asymptotic behavior of expanding direction)**.** *Let $\psi, u_{y,1}$ be given as in Theorem 3.1 and define $\rho := (\psi u_{y,1})^2$. The asymptotic behavior of the eigenpair $u^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\rho,0}(\Omega_L), \lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\rho,0}(\Omega_L)$ for $L \to \infty$ is*

$$\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\rho,0} = \frac{1}{L^2}\left(\nu^{(m)} + \mathcal{O}\left(\frac{1}{L}\right)\right), \tag{3.42}$$

$$L^{p/2} u^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\rho,0}(\boldsymbol{x}/L, \boldsymbol{y}) \rightharpoonup u^{(m)}_0(\boldsymbol{x}) \text{ weakly up to a subseq. in } H^1_{\mathcal{B}_d,\mathcal{B}_n}(\Omega_1), \tag{3.43}$$

where $(u_0^{(m)}, \nu^{(m)}) \in (H_0^1((0,1)^p) \setminus \{0\}) \times \mathbb{R}$ *is the solution to the p-dimensional homogenized eigenvalue problem*

$$\begin{cases} -\nabla \cdot \left(\bar{D} \nabla u_0^{(m)} \right) = \nu^{(m)} \bar{C} u_0^{(m)} \text{ in } (0,1)^p \\ u_0^{(m)} = 0 \text{ on } \partial(0,1)^p \end{cases}, \tag{3.44}$$

with the constant homogenized coefficients, $\bar{D} \in \mathbb{R}^{p \times p}$, $\bar{C} \in \mathbb{R}$, *given by*

$$\bar{D}_{ij} = \int_{(0,1)^p} \int_{(0,\ell)^q} \rho \left(\delta_{ij} + \frac{\partial \theta_j}{\partial x_i} \right) d\boldsymbol{y} \, d\boldsymbol{x}, \quad \bar{C} = \int_{(0,1)^p} \int_{(0,\ell)^q} \rho \, d\boldsymbol{y} \, d\boldsymbol{x}, \tag{3.45}$$

for $i, j = 1, \ldots, p$. *The corrector functions* $\{\theta_i(\boldsymbol{x}, \boldsymbol{y})\}_{1 \leq i \leq p}$ *are defined as cell problem solutions on the periodic unit cell, as*

$$\begin{cases} -\nabla \cdot \left(\rho(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \left(\boldsymbol{e}_i + \nabla \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \right) \right) = 0 \text{ in } \Omega_1 = (0,1)^p \times (0,\ell)^q \\ \tilde{\boldsymbol{x}} \mapsto \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \; \tilde{\boldsymbol{x}}\text{-periodic} \end{cases}. \tag{3.46}$$

Furthermore, it holds that $\nu^{(1)} < \nu^{(2)}$.

Proof. The proof is divided into five steps. We first apply a spatial transformation to identify the directional homogenization problem. In the second step, we show the existence of a weakly converging subsequence for the linear source problem. Then, the oscillating test function method provides the homogenized operators whose dimensions can be further reduced by considering the directional framework in the fourth step. The last step transfers the results to the eigenvalue problem.

Step 1: *Identification of a directional homogenization problem by transformation.*

To operate on fixed spatial domains, we map the problem from Ω_L to the reference domain $\Omega_1 = (0,1)^p \times (0,\ell)^q$ by $\boldsymbol{x} \mapsto \boldsymbol{x}/L =: \varepsilon \boldsymbol{x}$ and observe for the transformed weight function $\rho_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) := \rho(\boldsymbol{x}/\varepsilon, \boldsymbol{y})$ that $1/\rho_\varepsilon \in L^1_{\text{loc}}(\Omega_1)$ and $\rho_\varepsilon > 0$ a.e. in Ω_1. Thus, the correct framework is the weighted space $H^1(\Omega_1; \rho_\varepsilon)$. We now encode the Dirichlet boundary conditions on the $\boldsymbol{x}$-boundary (as in [216, p6]) in the sense of traces with $\Gamma_{\text{D}} := \{0,1\}^p \times (0,\ell)^q \subset \partial\Omega_1$ in the subspace

$$\mathbb{V}_{\rho_\varepsilon} = \{\phi \in H^1(\Omega_1; \rho_\varepsilon) \mid \phi = 0 \text{ on } \Gamma_{\text{D}}\} \subset H^1(\Omega_1; \rho_\varepsilon). \tag{3.47}$$

Here, $\mathbb{V}_{\rho_\varepsilon}$ is a Banach space since $\rho_\varepsilon = 0$ occurs only on the boundary Ω_1 (c.f. Remark 3.1). The weak form of the eigenvalue problem reads: Find $(u_\varepsilon^{(m)}, \lambda_\varepsilon^{(m)}) \in (\mathbb{V}_{\rho_\varepsilon} \setminus \{0\}) \times \mathbb{R}$, such that

$$\forall v \in \mathbb{V}_{\rho_\varepsilon} : \quad a_\varepsilon(u_\varepsilon^{(m)}, v) = \lambda_\varepsilon^{(m)} \int_{\Omega_1} \rho_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) u_\varepsilon^{(m)} v \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}, \tag{3.48}$$

with the bilinear form

$$a_\varepsilon(u_\varepsilon^{(m)}, v) = \int_{\Omega_1} \rho_\varepsilon(\boldsymbol{x}, \boldsymbol{y}) \left(\frac{\partial u_\varepsilon^{(m)}}{\partial x_i} \frac{\partial v}{\partial x_i} + \frac{1}{\varepsilon^2} \frac{\partial u_\varepsilon^{(m)}}{\partial y_i} \frac{\partial v}{\partial y_i} \right) \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y}, \tag{3.49}$$

using index notation. In Eq. (3.48), we moved the ε^2-scaling to $\lambda_\varepsilon^{(m)}$ as

$$\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\rho_\varepsilon,0} = \varepsilon^2 \lambda_\varepsilon^{(m)}, \tag{3.50}$$

which follows from the min-max principle. This operation will be justified later when the existence of this ε^2-transformed eigenvalue problem is shown for $\varepsilon \to 0$.

Step 2: *Extraction of a weakly converging subsequence for the linear equation.*

From [164, 165], we know that the homogenization of eigenvalue problems uses the same homogenized operators as for the corresponding source problem. Hence, we consider the bilinear form of the corresponding source problem to derive the homogenized operators. Therefore, given a family $f_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \in \mathbb{V}'_{\rho_\varepsilon}$ with $f_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \to f_0(\boldsymbol{x})$ in $\mathbb{V}'_{\rho_\varepsilon}$, we study the variational formulation

$$\forall v \in \mathbb{V}_{\rho_\varepsilon}: \quad a_\varepsilon(u_\varepsilon, v) = \langle f_\varepsilon, v\rangle_{\mathbb{V}'_{\rho_\varepsilon}\times\mathbb{V}_{\rho_\varepsilon}}. \tag{3.51}$$

The restriction of the family f_ε to $\boldsymbol{y}$-constant functions in the limit will be justified later when we show that the homogenized limit u_0 will have exactly this form. Thus, since we want to derive the eigenvalue problem from the source problem, f_ε has to mimic the properties of the sequence u_ε. The bilinear form $a_\varepsilon(u_\varepsilon, v)$ is $\mathbb{V}_{\rho_\varepsilon}$-elliptic (for $\varepsilon \leq 1$), since for all $u_\varepsilon \in \mathbb{V}_{\rho_\varepsilon}$, we have

$$\begin{aligned} a_\varepsilon(u_\varepsilon,u_\varepsilon) &\overset{\varepsilon\leq 1}{\geq} \|\nabla u_\varepsilon\|^2_{L^2(\Omega_1;\rho_\varepsilon)} \overset{\text{F.-in.}}{\geq} \frac{1}{2}\|\nabla u_\varepsilon\|^2_{L^2(\Omega_1;\rho_\varepsilon)} + \frac{\mathsf{C}_\mathrm{F}}{2}\|u_\varepsilon\|^2_{L^2(\Omega_1;\rho_\varepsilon)} \\ &\geq \frac{1}{2}\min\{1,\mathsf{C}_\mathrm{F}\}\|u_\varepsilon\|^2_{H^1(\Omega_1;\rho_\varepsilon)} =: \mathsf{C}\|u_\varepsilon\|^2_{H^1(\Omega_1;\rho_\varepsilon)}, \end{aligned} \tag{3.52}$$

after using the weighted Friedrichs inequality [185, p199] (for homogeneous Dirichlet boundary condition on parts of the boundary). Continuity also holds for all $\varepsilon > 0$ with a continuity constant proportional to $1/\varepsilon^2$. Thus, the problem is well-posed and admits a unique solution for all $\epsilon > 0$ in $\mathbb{V}_{\rho_\varepsilon}$ (Lax–Milgram, c.f. [78, p126]). We, however, are interested in precisely the limit $\varepsilon \to 0$, which, at first, seems to be problematic since the continuity constant would tend to infinity if we do not further specify the $\partial\boldsymbol{y}$-behavior in Eq. (3.49). However, we take (as in [216, p24]) u_ε in the bilinear form and use the coercivity to obtain

$$\mathsf{C}\|u_\varepsilon\|^2_{H^1(\Omega_1;\rho_\varepsilon)} \leq a_\varepsilon(u_\varepsilon,u_\varepsilon) = \langle f_\varepsilon, u_\varepsilon\rangle_{\mathbb{V}'_{\rho_\varepsilon}\times\mathbb{V}_{\rho_\varepsilon}} \leq \|f_\varepsilon\|_{H^{-1}(\Omega_1;\rho_\varepsilon)}\|u_\varepsilon\|_{H^1(\Omega_1;\rho_\varepsilon)}, \tag{3.53}$$

with the operator norm $\|f_\varepsilon\|_{H^{-1}(\Omega_1;\rho_\varepsilon)} = \|f_\varepsilon\|_{\mathbb{V}'_{\rho_\varepsilon}} \to \|f_0\|_{\mathbb{V}'_{\rho_\varepsilon}} \leq \mathsf{D}$ by our assumption on the family $f_\varepsilon \in \mathbb{V}'_{\rho_\varepsilon}$. Therefore, u_ε is uniformly bounded in $H^1(\Omega_1;\rho_\varepsilon)$ since $\|u_\varepsilon\|_{H^1(\Omega_1;\rho_\varepsilon)} \leq \frac{1}{\mathsf{C}}\|f_\varepsilon\|_{H^{-1}(\Omega_1;\rho_\varepsilon)} < \infty$. Now recall that $\rho_\varepsilon = \rho(\boldsymbol{x}/\varepsilon,\boldsymbol{y})$ is $\boldsymbol{x}$-periodic and thus weakly converges to its $\boldsymbol{x}$-average $\rho_0(\boldsymbol{y})$ [4, Lem. 1.8.]. From the boundedness of u_ε in $H^1(\Omega_1;\rho_\varepsilon)$, we can follow with [272, Prop. 2.1.] that there exists a $u_0 \in H^1(\Omega_1;\rho_0)$, such that there exists a converging subsequence of u_ε, still denoted by u_ε by abuse of notation, that weakly convergences in $H^1(\Omega_1;\rho_0)$. This ensures the

existence of the desired homogenized limit u_0 of u_ε as $\varepsilon \to 0$. We can also directly infer $\sqrt{\rho_\varepsilon}\frac{\partial u_\varepsilon}{\partial \boldsymbol{y}} \to \mathbf{0}$ in $L^2(\Omega_1)$ by taking the limit in

$$\begin{aligned}\frac{\mathsf{C}}{\varepsilon^2}\int_{\Omega_1}\rho_\varepsilon\left|\frac{\partial u_\varepsilon}{\partial \boldsymbol{y}}\right|^2 \mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y} &\leq a_\varepsilon(u_\varepsilon,u_\varepsilon)\\ &\leq \|f_\varepsilon\|_{H^{-1}(\Omega_1;\rho_\varepsilon)}\|u_\varepsilon\|_{H^1(\Omega_1;\rho_\varepsilon)} \leq \frac{1}{\mathsf{C}}\|f_\varepsilon\|^2_{H^{-1}(\Omega_1;\rho_\varepsilon)} < \infty,\end{aligned} \tag{3.54}$$

since the norm of u_ε is bounded by the norm of f_ε, which implies that

$$\lim_{\varepsilon\to 0}\left\|\sqrt{\rho_\varepsilon}\frac{\partial u_\varepsilon}{\partial \boldsymbol{y}} - \mathbf{0}\right\|_{L^2(\Omega_1)} = 0, \tag{3.55}$$

since $\frac{1}{C}\|f_\varepsilon\|^2_{H^{-1}(\Omega_1;\rho_\varepsilon)}$ is bounded for all ε, including $\varepsilon = 0$. Therefore, $\sqrt{\rho_\varepsilon}\frac{\partial u_\varepsilon}{\partial \boldsymbol{y}} \to \mathbf{0}$ in $L^2(\Omega_1)$, which will be important later to reduce the dimension of the homogenized equation from $p+q$ dimensions to just p. Since ρ_0 is nonzero a.e. on Ω_1 (recall that $\rho_\varepsilon = 0$ only happens on the $\boldsymbol{y}$-boundary), we have $\frac{\partial u_\varepsilon}{\partial \boldsymbol{y}} \to \mathbf{0}$ in $L^2(\Omega_1)$. Thus, a homogenized limit u_0 with $\partial u_0/\partial \boldsymbol{y} = \mathbf{0}$ exists for the sequence u_ε.

Step 3: *Derivation of the homogenized operators using oscillating test functions.*

Since we know that there exists a homogenized limit u_0, we aim to derive the corresponding homogenized equation for u_0. Therefore, consider

$$\xi_\varepsilon(\boldsymbol{x},\boldsymbol{y}) := \sqrt{\rho_\varepsilon(\boldsymbol{x},\boldsymbol{y})}\nabla u_\varepsilon(\boldsymbol{x},\boldsymbol{y}). \tag{3.56}$$

Following the usual arguments [216, p24], from the uniform boundedness of u_ε, it follows that

$$\|\xi_\varepsilon\|_{L^2(\Omega_1)} = |u_\varepsilon|_{H^1(\Omega_1;\rho_\varepsilon)} \leq \|u_\varepsilon\|_{H^1(\Omega_1;\rho_\varepsilon)} \leq \frac{1}{C}\|f_\varepsilon\|_{H^{-1}(\Omega_1;\rho_\varepsilon)} < \infty. \tag{3.57}$$

Therefore, we can again extract subsequences ξ_ε, still denoted by ξ_ε, such that $\xi_\varepsilon \rightharpoonup \xi_0$ in $L^2(\Omega_1)$ weakly. This convergence implies that the equation of interest

$$\langle \xi_\varepsilon, \nabla v\rangle_{L^2(\Omega_1)} = \langle f_\varepsilon, v\rangle_{\mathbb{V}_{\rho_\varepsilon}} \quad \forall v \in \mathbb{V}_{\rho_\varepsilon}, \tag{3.58}$$

has a limit for $\varepsilon \to 0$ as

$$\langle \xi_0, \nabla v\rangle_{L^2(\Omega_1)} = \langle f_0, v\rangle_{\mathbb{V}_{\rho_0}} \quad \forall v \in \mathbb{V}_{\rho_0}. \tag{3.59}$$

To explicitly state this limit equation, we need to calculate ξ_0. We employ the oscillatory test function method [4, p10] to overcome the problem of $\xi_\varepsilon = \sqrt{\rho(\boldsymbol{x}/\varepsilon,\boldsymbol{y})}\nabla u_\varepsilon$ being a product of two weakly converging functions and thus not simply being the product of both limits for $\varepsilon \to 0$. We need to adapt the method to account for the directional periodicity and the additional ε^{-2}-scaling of the $(\partial u_\varepsilon/\partial y_i)$-term in the bilinear form Eq. (3.49). Thus, let $\varphi \in \mathcal{D}((0,1)^p)$ be a smooth, only $\boldsymbol{x}$-dependent,

and compactly supported test function (i.e., $\varphi \in C_c^\infty((0,1)^p)$). Then, inspired by the first two terms in the asymptotic expansion for u_ε, we define the test function φ_ε as

$$\varphi_\varepsilon(\boldsymbol{x},\boldsymbol{y}) := \varphi(\boldsymbol{x}) + \varepsilon \sum_{i=1}^{p} \frac{\partial \varphi(\boldsymbol{x})}{\partial x_i} \theta_i(\boldsymbol{x}/\varepsilon, \boldsymbol{y}) \tag{3.60}$$

where, with $\tilde{\boldsymbol{x}} := \boldsymbol{x}/\varepsilon$, $\theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y})$ is the solution to the corrector problem Eq. (3.46), which admits for all $i = 1, \dots, p$ a unique solution $\theta_i \in H^1(\Omega_1; \rho_1) / \mathbb{R}$ (due to periodic boundary conditions). Since $\theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y})$ is $\tilde{\boldsymbol{x}}$-periodic, it converges weakly to its average in $H^1(\Omega_1; \rho_0)$ as $\varepsilon \to 0$. Thus, the expression $\varepsilon\theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y})$ in Eq. (3.60) converges to zero since $\varepsilon \to 0$. This implies that φ_ϵ has a well-defined limit $\varphi_\varepsilon \rightharpoonup \varphi_0 = \varphi(\boldsymbol{x})$ for $\varepsilon \to 0$.

In the following, we group the gradients into (p,q)-blocks by using the notation $\nabla(\cdot) := (\nabla_{\boldsymbol{x}}(\cdot), \nabla_{\boldsymbol{y}}(\cdot))^T$. As the derivative of φ_ε is required in the variational formulation, we derive from Eq. (3.60), using the chain and product rule, that

$$\nabla\varphi_\varepsilon = \sum_{i=1}^{p} \frac{\partial \varphi(\boldsymbol{x})}{\partial x_i} \left(e_i + \begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \\ \varepsilon \nabla_{\boldsymbol{y}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \end{pmatrix} \right) + \varepsilon \sum_{i=1}^{p} \left(\frac{\partial}{\partial x_i} \begin{pmatrix} \nabla_{\boldsymbol{x}} \varphi(\boldsymbol{x}) \\ 0 \end{pmatrix} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \right). \tag{3.61}$$

We then insert the test function φ_ε into the bilinear form Eq. (3.49) to obtain

$$\begin{aligned} a_\varepsilon(u_\varepsilon, \varphi_\varepsilon) = &\int_{\Omega_1} \rho_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \nabla u_\varepsilon \cdot \left(\sum_{i=1}^{p} \frac{\partial \varphi(\boldsymbol{x})}{\partial x_i} \left(e_i + \begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \\ \varepsilon^{-1} \nabla_{\boldsymbol{y}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \end{pmatrix} \right) \right) \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y} \\ &+ \varepsilon \int_{\Omega_1} \rho_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \nabla u_\varepsilon \cdot \left(\sum_{i=1}^{p} \left(\frac{\partial}{\partial x_i} \begin{pmatrix} \nabla_{\boldsymbol{x}} \varphi(\boldsymbol{x}) \\ 0 \end{pmatrix} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \right) \right) \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}. \end{aligned} \tag{3.62}$$

The last term in Eq. (3.62) vanishes in the limit since it can be bounded by a constant times ε by the Cauchy–Schwarz inequality as the (φ, θ_i)-term is uniformly bounded in $L^2(\Omega_1; \rho_\varepsilon)$ by uniformly boundedness of the data in Eq. (3.46) and φ-smoothness. The other term, ∇u_ε, is uniformly bounded in $L^2(\Omega_1; \rho_\varepsilon)$ by Eq. (3.57). Integration by parts (with Dirichlet in $\boldsymbol{x}$- and trivially fulfilled Neumann data in $\boldsymbol{y}$-direction) in the other term of Eq. (3.62) yields

$$\begin{aligned} &\int_{\Omega_1} \rho_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \nabla u_\varepsilon \cdot \left(\sum_{i=1}^{p} \frac{\partial \varphi(\boldsymbol{x})}{\partial x_i} \left(e_i + \begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \\ \varepsilon^{-1} \nabla_{\boldsymbol{y}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \end{pmatrix} \right) \right) \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y} \\ &= -\int_{\Omega_1} u_\varepsilon \nabla \cdot \left(\rho_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \sum_{i=1}^{p} \frac{\partial \varphi(\boldsymbol{x})}{\partial x_i} \left(e_i + \begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \\ \varepsilon^{-1} \nabla_{\boldsymbol{y}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \end{pmatrix} \right) \right) \mathrm{d}\boldsymbol{x}\, \mathrm{d}\boldsymbol{y}. \end{aligned} \tag{3.63}$$

The divergence term in Eq. (3.63) can be further simplified to

$$\begin{aligned} &\nabla \cdot \left(\rho_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \sum_{i=1}^{p} \frac{\partial \varphi(\boldsymbol{x})}{\partial x_i} \left(e_i + \begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \\ \varepsilon^{-1} \nabla_{\boldsymbol{y}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \end{pmatrix} \right) \right) \\ &= \sum_{i=1}^{p} \frac{\partial}{\partial x_i} \begin{pmatrix} \nabla_{\boldsymbol{x}} \varphi(\boldsymbol{x}) \\ 0 \end{pmatrix} \cdot \rho_\varepsilon(\boldsymbol{x},\boldsymbol{y}) \left(e_i + \begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \\ 0 \end{pmatrix} \right) \\ &+ \varepsilon^{-1} \sum_{i=1}^{p} \frac{\partial \varphi(\boldsymbol{x})}{\partial x_i} \left[\begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \\ \nabla_{\boldsymbol{y}} \end{pmatrix} \cdot \left(\rho(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \left(e_i + \begin{pmatrix} \nabla_{\tilde{\boldsymbol{x}}} \\ \nabla_{\boldsymbol{y}} \end{pmatrix} \theta_i(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \right) \right) \right], \end{aligned} \tag{3.64}$$

From Eq. (3.64), we extract

$$\varepsilon^{-1}\begin{pmatrix}\nabla_{\tilde{\boldsymbol{x}}}\\ \nabla_{\boldsymbol{y}}\end{pmatrix}\cdot\left(\rho(\tilde{\boldsymbol{x}},\boldsymbol{y})\left(e_i+\begin{pmatrix}\nabla_{\tilde{\boldsymbol{x}}}\\ \nabla_{\boldsymbol{y}}\end{pmatrix}\theta_i(\tilde{\boldsymbol{x}},\boldsymbol{y})\right)\right), \tag{3.65}$$

which is zero (even for $\varepsilon\to 0$) due to the particular definition of the correctors θ_i in Eq. (3.46). Here, we notice that the ε^{-2}-scaling in the $\boldsymbol{y}$-term of the initial expression precisely aligns with the extruding additional ε^{-1} that appeared in Eq. (3.65) by the chain rule. Thus, the only remaining term of Eq. (3.64) is

$$\sum_{i=1}^{p}\frac{\partial}{\partial x_i}\begin{pmatrix}\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x})\\ 0\end{pmatrix}\cdot\rho_\varepsilon(\boldsymbol{x},\boldsymbol{y})\left(e_i+\begin{pmatrix}\nabla_{\tilde{\boldsymbol{x}}}\theta_i(\tilde{\boldsymbol{x}},\boldsymbol{y})\\ 0\end{pmatrix}\right), \tag{3.66}$$

which is bounded in $L^2(\Omega_1)$ and, thus, weakly converges as $\varepsilon\to 0$ to its average in the $\tilde{\boldsymbol{x}}$-direction [4, Lem. 1.8.].

In Eq. (3.63), now recall that u_ε converges strongly to u_0 in $L^2(\Omega_1;\rho_0)$ (by the Rellich theorem, c.f. [5, Thm. 4.3.21]). Thus, we can take the limit of the right-hand side in Eq. (3.63). So in total, we can take the limit of Eq. (3.62), which is the product of the limit of $u_\varepsilon\to u_0$ (strongly) with the weak limit Eq. (3.66) of the divergence term. Thus the weak form Eq. (3.51) reduces for $\varepsilon\to 0$ to

$$\begin{aligned}
&-\int_{\Omega_1}u_0(\boldsymbol{x})\nabla\cdot\left(\int_{(0,1)^p}\rho(\tilde{\boldsymbol{x}},\boldsymbol{y})\sum_{i=1}^{p}\frac{\partial\varphi(\boldsymbol{x})}{\partial x_i}\left(e_i+\begin{pmatrix}\nabla_{\tilde{\boldsymbol{x}}}\theta_i(\tilde{\boldsymbol{x}},\boldsymbol{y})\\ 0\end{pmatrix}\right)\mathrm{d}\tilde{\boldsymbol{x}}\right)\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\\
&=\lim_{\varepsilon\to 0}\int_{\Omega_1}\rho_\varepsilon(\boldsymbol{x},\boldsymbol{y})\nabla u_\varepsilon(\boldsymbol{x},\boldsymbol{y})\cdot\nabla\varphi_\varepsilon(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\\
&=\lim_{\varepsilon\to 0}\int_{\Omega_1}\rho_\varepsilon(\boldsymbol{x},\boldsymbol{y})f_\varepsilon(\boldsymbol{x},\boldsymbol{y})\varphi_\varepsilon(\boldsymbol{x},\boldsymbol{y})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\\
&=\int_{\Omega_1}\rho_0(\boldsymbol{y})f_0(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}.
\end{aligned} \tag{3.67}$$

We can rewrite the left-hand side of Eq. (3.67) using a compact notation as

$$\begin{aligned}
&\left[\int_{(0,1)^p}\rho(\tilde{\boldsymbol{x}},\boldsymbol{y})\sum_{i=1}^{p}\frac{\partial\varphi(\boldsymbol{x})}{\partial x_i}\left(e_i+\begin{pmatrix}\nabla_{\tilde{\boldsymbol{x}}}\theta_i(\tilde{\boldsymbol{x}},\boldsymbol{y})\\ 0\end{pmatrix}\right)\mathrm{d}\tilde{\boldsymbol{x}}\right]_j\\
&=\sum_{i=1}^{p}\frac{\partial\varphi(\boldsymbol{x})}{\partial x_i}\int_{(0,1)^p}\rho(\tilde{\boldsymbol{x}},\boldsymbol{y})\left(\delta_{ij}+\frac{\partial\theta_i}{\partial\tilde{x}_j}\right)\mathrm{d}\tilde{\boldsymbol{x}}=:\sum_{i=1}^{p}\frac{\partial\varphi(\boldsymbol{x})}{\partial x_i}\tilde{D}_{ji}(\boldsymbol{y}),
\end{aligned} \tag{3.68}$$

where we identify the last expression as $[\tilde{D}^T(\boldsymbol{y})\nabla\varphi(\boldsymbol{x})]_j$. As the last step, we reverse the integration by parts and obtain the variational formulation of the homogenized equation for u_0, which is still posed on the $(p+q)$-dimensional domain Ω_1, but with u_0 only $\boldsymbol{x}$-dependent according to Eq. (3.55). The problem then reads: Find $u_0(\boldsymbol{x})\in\mathbb{V}_{\rho_0}$, such that

$$\int_{\Omega_1}\tilde{D}(\boldsymbol{y})\nabla u_0(\boldsymbol{x})\cdot\nabla\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}=\int_{\Omega_1}\tilde{C}(\boldsymbol{y})f_0(\boldsymbol{x})\varphi(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\,\mathrm{d}\boldsymbol{y}\quad\forall\varphi\in C_c^\infty((0,1)^p), \tag{3.69}$$

with the $\boldsymbol{y}$-dependent operators

$$\tilde{D}_{ij}(\boldsymbol{y}) = \int_{(0,1)^p} \rho(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \left(\delta_{ij} + \frac{\partial \theta_j}{\partial \tilde{x}_i}\right) \mathrm{d}\tilde{\boldsymbol{x}}, \quad \tilde{C}(\boldsymbol{y}) = \int_{(0,1)^p} \rho(\tilde{\boldsymbol{x}}, \boldsymbol{y}) \, \mathrm{d}\tilde{\boldsymbol{x}}, \tag{3.70}$$

for $i, j = 1, \ldots, p$. In Eq. (3.70), we remark that these operators look very similar to the usual homogenized operators, e.g., in [8], with the difference that the integration only takes place in the p expanding directions over $(0,1)^p$.

Step 4: *Dimension reduction of the homogenized linear equation.*

In our setup, we can, however, further reduce the homogenized limit equation Eq. (3.69) since, by definition, $\nabla\varphi(\boldsymbol{x}) = (\nabla_{\boldsymbol{x}}\varphi(\boldsymbol{x}), 0)^T$. This allows us to concretize further that $u_0(\boldsymbol{x}) \in H_0^1((0,1)^p)$ since $u_0(\boldsymbol{x}) \in \mathbb{V}_{\rho_0(\boldsymbol{y})}$ implies $u_0(\boldsymbol{x}) = 0$ on $\partial(0,1)^p$ and $\|u_0(\boldsymbol{x})\|_{H^1((0,1)^p)} < \infty$ since for any $u(\boldsymbol{x}) \in H^1(\Omega_1; \rho_0)$ with $\rho_0(\boldsymbol{y}) > 0$ a.e. in $(0,\ell)^q$, it holds that

$$\|u(\boldsymbol{x})\|^2_{H^1(\Omega_1;\rho_0)} = \left(\int_{(0,1)^q} \rho_0(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{y}\right) \|u(\boldsymbol{x})\|^2_{H^1((0,1)^p)} = \overline{\rho_0}^y \|u(\boldsymbol{x})\|^2_{H^1((0,1)^p)} < \infty, \tag{3.71}$$

since $\overline{\rho_0}^y \in \mathbb{R}$ is a strictly positive constant. Thus, the homogenized equation reduces from the $(p+q)$- to the p-dimensional variational problem: Find $u_0 \in H_0^1((0,1)^p)$, such that

$$\int_{(0,1)^p} \bar{D}\nabla u_0(\boldsymbol{x}) \cdot \nabla\varphi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} = \int_{(0,1)^p} \bar{C} f_0(\boldsymbol{x})\varphi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \quad \forall \varphi \in C_c^\infty((0,1)^p), \tag{3.72}$$

with the constant homogenized coefficients, defined by Eq. (3.45) as the integral of $\tilde{C}(\boldsymbol{y})$ and $\tilde{D}(\boldsymbol{y})$ from Eq. (3.70) over $(0,1)^q$.

The homogenized equation Eq. (3.72) is formulated on $H_0^1((0,1)^p)$. Recall that the test function $\varphi \in C_c^\infty((0,1)^p)$ was chosen arbitrarily. Since $C_c^\infty((0,1)^p)$ is dense in $H_0^1((0,1)^p)$ by the definition of H_0^1 as the closure of C_c^∞ under the H^1-norm [5, Def. 4.3.8.], Eq. (3.69) holds $\forall\varphi \in H_0^1((0,1)^p)$. As the homogenized operator satisfies coercivity (c.f. [216, Rem. 2.6.]), the theorem of Lax–Milgram ensures the uniqueness of the homogenized limit u_0. This, on the other hand, implies that any subsequence of u_ε converges to u_0 in the limit. Thus, the entire sequence u_ε converges to the same limit u_0 following the standard arguments from, e.g., [4].

Step 5: *Derivation of the homogenized eigenvalue equation.*

Since we now have derived the homogenized equation for the source problem, we can directly deduce from [164, Thm. 2.1.] that the eigenvalues and -functions converge to the homogenized eigenvalue equation, posed with the same operator as in Eq. (3.72), resulting in

$$(\lambda_\varepsilon^{(m)}, u_\varepsilon^{(m)}) \to (\nu^{(m)}, u_0^{(m)}) \text{ in } \mathbb{R} \times \left(H_0^1((0,1)^p) \text{ weakly up to subseq.}\right), \tag{3.73}$$

where the homogenized eigenpair is defined through Eq. (3.44). We furthermore have $\lambda_\varepsilon^{(m)} = \nu^{(m)} + \mathcal{O}(\varepsilon)$ [165, p201] [230, p1638] [10, p942]. The convergence of the eigenfunctions holds up to a subsequence because of the eigenvalue multiplicity of

the homogenized limit. To account for the normalization constraint after the initial transformation of $\boldsymbol{x} \mapsto \varepsilon\boldsymbol{x}$, we note that $\|u_0(\cdot,\cdot)\|_{L^2(\Omega_L)} = L^{p/2}\|u_0(\cdot/L,\cdot)\|_{L^2(\Omega_1)}$ by the transformation rule and recall the $(1/L^2)$-scaling from Eq. (3.50) for the eigenvalues, which implies Eq. (3.42).

The limit eigenvalue $\nu^{(m)}$ is simple and $\nu^{(1)} < \nu^{(2)} < \nu^{(3)} < \cdots \to \infty$ by the Sturm–Liouville theory for the particular case of $p = 1$ with $\bar{D}_{11}, \bar{C} > 0$. Furthermore, we have $\nu^{(1)} < \nu^{(2)} \leq \nu^{(3)} \leq \cdots \to \infty$ for the general case of $p \geq 2$ since the eigenvalue problem is elliptic. However, multiplicities could exceed one for higher eigenvalues. □

We are now ready to prove the quasi-optimality of the spectral shift $\sigma = \lambda_{\varphi_y}$:

Theorem 3.3. *For the quasi-optimal shift $\sigma = \lambda_{\varphi_y} = \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,1,V}(\Omega_1)$, the asymptotic shifted fundamental eigenvalue ratio of the linear periodic Schrödinger eigenvalue problem Eq. (3.2) converges to a positive constant $\mathsf{C} < 1$ as $L \to \infty$, that is*

$$0 \leq \frac{\lambda^{(1)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L) - \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,1,V}(\Omega_1)}{\lambda^{(2)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L) - \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,1,V}(\Omega_1)} = \frac{\lambda^{(1)}_{\mathcal{B}_d,\mathcal{B}_n,\varphi_y^2,0}(\Omega_L)}{\lambda^{(2)}_{\mathcal{B}_d,\mathcal{B}_n,\varphi_y^2,0}(\Omega_L)} \to \mathsf{C} < 1. \tag{3.74}$$

Proof. The proof follows from Theorem 3.2 since $L^2\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\varphi_y^2,0}(\Omega_L) = \nu^{(m)} + o\left(\frac{1}{L}\right)$ and $\nu^{(1)} < \nu^{(2)}$. □

We will see later that pre-asymptotic effects lead to a non-monotonic convergence of Eq. (3.74). However, since the convergence holds in the limit, we can make a statement for uniform boundedness if L is sufficiently large.

Corollary 3.1. *There exists a constant $\mathsf{D} \in [0,1)$ and a length $L^* \in \mathbb{R}^+$, such that the quasi-optimally shifted ratio from Theorem 3.3 is uniformly bounded from above by D for all $L > L^*$. That is*

$$0 \leq \frac{\lambda^{(1)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L) - \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,1,V}(\Omega_1)}{\lambda^{(2)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L) - \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,1,V}(\Omega_1)} < \mathsf{D} < 1 \quad \forall L > L^*. \tag{3.75}$$

Proof. The proof follows directly from the convergence result of Theorem 3.3. □

Remark 3.4. The quasi-optimal shift $\sigma = \lambda^{(1)}_{\mathcal{B}_d,\mathcal{B}_d,1,V}(\Omega_L)$ does not affect the absolute eigenvalue ordering in the sense that $0 < |\lambda_L^{(1)} - \sigma| < |\lambda_L^{(2)} - \sigma| \leq |\lambda_L^{(3)} - \sigma| \leq \cdots \to \infty$ since all $\lambda_L^{(i)}$ are positive and $\sigma < \lambda_L^{(1)}$. This property ensures, for example, that the unshifted and the σ-shifted inverse power method (see Definition 3.2) always converge to the same eigenpair.

Theorem 3.2 gives an abstract description of the homogenized equation. However, for our present setup, we can even solve the equation analytically (which will be important later in Section 3.4.1):

Remark 3.5. The homogenization problem in Theorem 3.2 is posed with an isotropic operator ρI, and ρ is either periodic or zero on the unit cell boundaries $\partial\Omega_1$. Thus, every column of ρI is a solenoidal vector field in Ω_1 in the integral sense by the divergence theorem. Hence, we can conclude with [158, p17] that the homogenized operator is diagonal. Then, the diagonality allows us to explicitly state the homogenized eigenpairs as the Laplacian eigenfunctions on the hyper rectangle with scaled Laplacian eigenvalues as $\nu^{(m)} = \pi^2 \left(\sum_{i=1}^{p} \bar{D}_{ii} m_i^2\right) / \bar{C}$ and $u^{(m)}(\boldsymbol{x}) = \mathcal{N}^{(m)} \prod_i^p \sin(m_i \pi x_i)$, where the set $\mathcal{M} = \{m_i, \ldots, m_p\} \in \mathbb{N}^p, |\mathcal{M}| = m$, is chosen to minimize $\nu^{(m)}$. The $\mathcal{N}^{(m)}$ factors are defined by the normalization condition $\int_{(0,1)^p} \bar{C} \left(u^{(m)}\right)^2 = 1$.

We now return to the convergence properties of the eigenvalue solvers. Theorem 3.3 implies a constant number of iterations for all eigensolvers that are shift-and-invert preconditioned with $\sigma = \lambda^{(1)}_{\mathcal{B}_\#, \mathcal{B}_d, 1, V}(\Omega_1)$ and depend on the fundamental ratio. With this strategy, the eigensolver can reach a given residual norm with a constant number of iterations for all $L \to \infty$.

3.3 Spatial Discretization and Iterative Eigensolvers

To solve the eigenvalue problem Eq. (3.2) numerically, we will discretize the continuous equation on a finite-dimensional space. Then, we solve the resulting system with a preconditioned algebraic eigensolver.

3.3.1 Galerkin Finite Element Approach

Consider a conforming and shape-regular partition $\mathcal{T}_h$ of the domain Ω_L into finite elements $\tau \in \mathcal{T}_h$, which have a polygonal shape. We write $\mathcal{T}_h$ for partitions where every element has a diameter of at most $2h$ [240, p36]. Define the finite element subspace $\mathbb{H}_h(\Omega_L) \subset H_0^1(\Omega_L)$, consisting of polynomial functions with total degree r from the polynomial space $\mathcal{P}_r$, to be $\mathbb{H}_h(\Omega_L) = \{u \in H_0^1(\Omega_L) \mid u|_\tau \in \mathcal{P}_r(\tau)\ \forall \tau \in \mathcal{T}_h\}$. We then search for a discrete solution $\phi_h^{(m)} \in (\mathbb{H}_h(\Omega_L) \setminus \{0\})$, such that

$$\forall v_h \in \mathbb{H}_h(\Omega_L): \quad \int_{\Omega_L} \nabla \phi_h^{(m)} \cdot \nabla v_h \, \mathrm{d}\boldsymbol{z} + \int_{\Omega_L} V \phi_h^{(m)} v_h \, \mathrm{d}\boldsymbol{z} = \lambda_h^{(m)} \int_{\Omega_L} \phi_h^{(m)} v_h \, \mathrm{d}\boldsymbol{z}. \tag{3.76}$$

Let now $\boldsymbol{x}_h^{(m)}$ be the coefficient vector that represents $\phi_h^{(m)}$ in a given basis of $\mathbb{H}_h(\Omega_L)$. We then obtain the equivalent generalized algebraic eigenvalue problem: Find $\boldsymbol{x}_h^{(m)} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$, such that

$$\boldsymbol{A}\boldsymbol{x}_h^{(m)} = \lambda_h^{(m)} \boldsymbol{B}\boldsymbol{x}_h^{(m)}, \tag{3.77}$$

where $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ consists of the usual stiffness matrix plus the contribution from the potential, and $\boldsymbol{B} \in \mathbb{R}^{n \times n}$ denotes the mass matrix. Both $\boldsymbol{A}$ and $\boldsymbol{B}$ as finite representations of the continuous operators in Eq. (3.2) are symmetric positive definite. Since the discrete problem is formulated on a subspace $\mathbb{H}_h(\Omega_L) \subset H_0^1(\Omega_L)$, we have by

the min-max characterization that $\lambda^{(m)} \leq \lambda_h^{(m)}$. Furthermore, we have $\lambda_h^{(m)} \to \lambda^{(m)}$ for $h \to 0$ [28, 240].

For the calculation of the quasi-optimal shift λ_{φ_y}, we solve

$$\forall v_h \in \mathbb{H}_h^{\varphi_y}(\Omega_1): \quad \int_{\Omega_1} \nabla \varphi_{y,h}^{(1)} \cdot \nabla v_h \, \mathrm{d}\boldsymbol{z} + \int_{\Omega_1} V \varphi_{y,h}^{(1)} v_h \, \mathrm{d}\boldsymbol{z} = \lambda_{\varphi_y,h}^{(1)} \int_{\Omega_1} \varphi_{y,h}^{(1)} v_h \, \mathrm{d}\boldsymbol{z}, \tag{3.78}$$

where $\mathbb{H}_h^{\varphi_y}(\Omega_1) := \left\{ u \in H^1_{\mathcal{B}_\#,\mathcal{B}_d}(\Omega_1) \;\middle|\; u|_\tau \in \mathcal{P}_r(\tau) \; \forall \tau \in (\mathcal{T}_h \cap \Omega_1) \right\}$ with the same $\mathcal{T}_h$ and r as for $\mathbb{H}_h(\Omega_L)$ (assuming Ω_1-aligned elements).

3.3.2 Quasi-Optimally Preconditioned Eigenvalue Algorithms

To solve the resulting discrete eigenvalue problem Eq. (3.77), we use the analytic results from Section 3.2 to obtain the quasi-optimal shift as

$$\sigma = \lambda_\infty = \lim_{L\to\infty} \lambda_L^{(1)} = \lambda_{\varphi_y} \approx \lambda_{\varphi_y,h}^{(1)}, \tag{3.79}$$

by combining the results of Theorems 3.1 and 3.2. Using σ, we construct the preconditioner $\boldsymbol{P} = (\boldsymbol{A} - \sigma \boldsymbol{B})^{-1}$. With the generalized Rayleigh quotient $R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}) = (\boldsymbol{x}^T \boldsymbol{A} \boldsymbol{x})/(\boldsymbol{x}^T \boldsymbol{B} \boldsymbol{x})$, we define:

Definition 3.2 (Shifted Inverse Power Method, abbreviated by IP_σ)**.** Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ and a start vector $\boldsymbol{x}_0 \in \mathbb{R}^n$ be given, repeat

$$\tilde{\boldsymbol{x}}_k = \boldsymbol{P}\boldsymbol{B}\boldsymbol{x}_{k-1}, \quad \boldsymbol{x}_k = \tilde{\boldsymbol{x}}_k / \sqrt{\tilde{\boldsymbol{x}}_k^T \boldsymbol{B} \tilde{\boldsymbol{x}}_k}, \quad \lambda_k = R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_k), \tag{3.80}$$

until $\|\boldsymbol{A}\boldsymbol{x}_k - \lambda_k \boldsymbol{B}\boldsymbol{x}_k\|_2 < \texttt{TOL}$ or $k > k_{\max}$.

Definition 3.3 (Locally Optimal Preconditioned Conjugate Gradient Method, abbreviated by LOPCG_σ)**.** Let $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$ and the start vectors $\boldsymbol{x}_{-1}, \boldsymbol{x}_0 \in \mathbb{R}^n$ be given, repeat

$$\begin{aligned} \boldsymbol{w}_k &= \boldsymbol{P}(\boldsymbol{A}\boldsymbol{x}_{k-1} - R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_{k-1})\boldsymbol{B}\boldsymbol{x}_{k-1}), \quad S_k = \operatorname{span}(\{\boldsymbol{x}_{k-1}, \boldsymbol{w}_k, \boldsymbol{x}_{k-2}\}) \\ \tilde{\boldsymbol{x}}_k &= \arg\min_{\boldsymbol{y}\in S_k} R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{y}), \quad \boldsymbol{x}_k = \tilde{\boldsymbol{x}}_k / \sqrt{\tilde{\boldsymbol{x}}_k^T \boldsymbol{B} \tilde{\boldsymbol{x}}_k}, \quad \lambda_k = R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_k), \end{aligned} \tag{3.81}$$

until $\|\boldsymbol{A}\boldsymbol{x}_k - \lambda_k \boldsymbol{B}\boldsymbol{x}_k\|_2 < \texttt{TOL}$ or $k > k_{\max}$.

In Eq. (3.81), the locally optimal step is calculated by minimizing in a 3-dimensional subspace with the standard Rayleigh–Ritz method [31] as $\tilde{\boldsymbol{x}}_k = \alpha_1 \boldsymbol{x}_{k-1} + \alpha_2 \boldsymbol{w}_k + \alpha_3 \boldsymbol{x}_{k-2}$, where the coefficients $\boldsymbol{\alpha} \in \mathbb{R}^3$ are derived from the smallest eigenpair solution of the 3-dimensional eigenvalue problem $\boldsymbol{V}^T\boldsymbol{A}\boldsymbol{V}\boldsymbol{\alpha} = \lambda^{(1)}\boldsymbol{V}^T\boldsymbol{B}\boldsymbol{V}\boldsymbol{\alpha}$ with $\boldsymbol{V} = \begin{bmatrix} \boldsymbol{x}_{k-1} & \boldsymbol{w}_k & \boldsymbol{x}_{k-2} \end{bmatrix} \in \mathbb{R}^{n\times 3}$.

The above two methods represent the class of gap-dependent iterative eigenvalue algorithms. Optimization-inspired Riemannian gradient algorithms also depend on the fundamental ratio [146, Thm. 3.2]. Alternative approaches, such as the Rayleigh quotient iteration or block algorithms, are not considered in our setup since the former has no guaranteed convergence to the ground state [31, p53]. At the same time, the latter requires an L-proportional block size to retain a quasi-optimal convergence [31, p54].

3.4 Numerical Experiments

This section concerns the numerical evaluation of the proposed eigensolver preconditioner. We implemented our method using the `Gridap` [30] framework in the Julia programming language [44]. `Gridap` turned out to be a very well-suited framework for our tests since it allowed us to quickly implement weak formulations in a high-level fashion, similar to the `FEniCS` [12, 250] framework in Python. For reproducibility, we provide all examples publicly in [248].

3.4.1 Homogenization of a Degenerate Eigenvalue Problem With Two Expanding Directions in Three Dimensions

Before we employ the constructed preconditioner for the linear Schrödinger eigenvalue problem Eq. (3.2), we first investigate the homogenization results of Theorem 3.2 since these results can be applied and studied independently. Thus, the theoretical predictions about the convergence of the m-dependent contribution $u^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\rho,0}(\Omega_L), \lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_n,\rho,0}(\Omega_L)$ in three dimensions $(p = 2, q = 1)$ are studied numerically. We prescribe the weight function by

$$\rho(\boldsymbol{x},\boldsymbol{y}) = \left(\tfrac{27}{4} y_1^2(1-y_1)\left(10\cos(\pi x_1)^2 + 10\cos(\pi x_2)^2 + \tfrac{11}{10} - \sin(\pi y_1)^2\right)\right)^2. \quad (3.82)$$

Note that we do not set $\rho = (\psi u_{y,1})$ as in Theorem 3.2 since we want to demonstrate the results for the more general case of ρ not being induced by eigenfunctions but only satisfying the periodicity- and zero-condition on the $\boldsymbol{x}$- and $\boldsymbol{y}$-boundary respectively. The weight function ρ in Eq. (3.82) is positive a.e. and vanishes only on the $\boldsymbol{y}$-boundary. By construction, ρ is $\boldsymbol{x}$-periodic, thus fulfilling all requirements of Theorem 3.2. We intentionally use the $\boldsymbol{x}$-symmetry also to confirm the convergence of degenerate eigenpairs. For a better evaluation, we do not solve for Ω_L but solve an equivalent problem on the reference domain Ω_1, where we factorized the $(1/L^2)$-scaling (see Theorem 3.2) of the eigenvalue without affecting the eigenfunctions. To be precise, we check if the solution to

$$\begin{cases} -\nabla\cdot\left(\rho(L\boldsymbol{x},\boldsymbol{y})\,\mathrm{diag}\left(1,1,\tfrac{1}{L^2}\right)\nabla u^{(m)}_{1/L}\right) = \lambda^{(m)}_{1/L}\rho(L\boldsymbol{x},\boldsymbol{y})u^{(m)}_{1/L} \text{ in } (0,1)^3 \\ u^{(m)}_{1/L} = 0 \text{ on } \partial(0,1)^2\times(0,1) \end{cases}, \quad (3.83)$$

converge to $(u_0^{(m)}, \nu^{(m)})$ from Eq. (3.44) in the limit for $L\to\infty$. The calculation of this homogenized limit first needs the corrector functions to define the homogenized operators. Thus, we solve the corrector equation Eq. (3.46) using $\mathbb{Q}_2$ finite elements on a structured mesh with 300 intervals per direction. These corrector solutions allow the construction of the homogenized coefficients (according to Eqs. (3.44) and (3.45)) with $\bar{D} \approx \mathrm{diag}(38.75893, 38.75893)$ and $\bar{C} \approx 57.86864$. We observe that $\bar{D}_{11} = \bar{D}_{22}$ as the result of choosing an (x_1,x_2)-symmetric weight function ρ. The homogenized diffusion matrix is diagonal since we have $\int_{\Omega_1}\nabla\cdot(\rho I) = 0$ by the divergence theorem as ρ defined by Eq. (3.82) is either periodic or zero on the boundary of the unit cube,

which resembles the case of Remark 3.5. Therefore, we can solve the homogenized equation analytically (with the expressions from Remark 3.5) to obtain

$$\begin{gathered}\nu^{(1)} = \frac{\pi^2\left(1^2\bar{D}_{11} + 1^2\bar{D}_{22}\right)}{\bar{C}} = \frac{2\pi^2\bar{D}_{11}}{\bar{C}},\ \nu^{(2)} = \nu^{(3)} = \frac{5\pi^2\bar{D}_{11}}{\bar{C}},\ \nu^{(4)} = \frac{8\pi^2\bar{D}_{11}}{\bar{C}} \\ u_0^{(1)} = \mathcal{N}\sin(\pi x_1)\sin(\pi x_2),\ u_0^{(2)} = \mathcal{N}\sin(2\pi x_1)\sin(\pi x_2), \\ u_0^{(3)} = \mathcal{N}\sin(\pi x_1)\sin(2\pi x_2),\ u_0^{(4)} = \mathcal{N}\sin(2\pi x_1)\sin(2\pi x_2)\end{gathered} \tag{3.84}$$

with the normalization constant $\mathcal{N} = 2/\sqrt{\bar{C}} \approx 0.26291$ since

$$\left(\int_0^1 \sin^2(m_1\pi x_1)\,\mathrm{d}x_1\right)\cdots\left(\int_0^1 \sin^2(m_p\pi x_p)\,\mathrm{d}x_p\right) = 2^{-p/2}\ \forall \boldsymbol{m}\in\mathbb{N}^p. \tag{3.85}$$

We then solve the eigenvalue problem Eq. (3.83) using the Galerkin method with $\mathbb{Q}_2$ elements and a structured partition of both expanding directions into $12L$ intervals. According to Theorem 3.2, the non-relevant third direction is only discretized with six partitions since it is not relevant in the limit. Finally, we solve the corresponding algebraic eigenvalue problem using a block LOPCG method up to a tolerance of 10^{-6}.

For the error comparison, we project the analytical solutions Eq. (3.84) into the corresponding subspace $\mathbb{H}_h$. Since $\nu^{(2)} = \nu^{(3)}$ by our construction of ρ, the corresponding eigenspace is two-dimensional, and the eigensolver returns some basis of this space. To resolve these spatial rotations and allow for an error comparison, we align the second and third eigenfunction by modifying their discrete eigenvectors with

$$\boldsymbol{x}_h^{(2)} = \langle \boldsymbol{x}_h^{(2)}, \boldsymbol{x}_0^{(2)}\rangle \boldsymbol{x}_h^{(2)} + \langle \boldsymbol{x}_h^{(3)}, \boldsymbol{x}_0^{(2)}\rangle \boldsymbol{x}_h^{(3)},\quad \boldsymbol{x}_h^{(3)} = \langle \boldsymbol{x}_h^{(2)}, \boldsymbol{x}_0^{(3)}\rangle \boldsymbol{x}_h^{(2)} + \langle \boldsymbol{x}_h^{(3)}, \boldsymbol{x}_0^{(3)}\rangle \boldsymbol{x}_h^{(3)}, \tag{3.86}$$

where $\boldsymbol{x}_0^{(m)}$ denotes the m-th homogenized eigenvector. The resulting discrete eigenfunctions $u_{1/L,h}^{(m)}$ for $m = 1,2,3,4$ are visualized in Fig. 3.4 for $L \in \{2^0,\dots,2^4\}$ together with the corresponding homogenized solutions $u_0^{(m)}$. We can observe that for larger domain lengths L, the eigenfunctions converge to their corresponding limits if we would neglect the oscillatory isolines that indicate strong gradients. This observation corresponds to our theoretical results that the convergence is only weak when considering the $H^1(\Omega_1)$-norm. To quantify the convergence, we evaluate the relative $L^2(\Omega_1)$-error of the eigenfunctions and the relative eigenvalue error in Fig. 3.5 for $L \in \mathbb{R}$ with a sampling rate of $\Delta L = 0.1$. We measure a first-order converge for the L^2-error and at least first-order convergence for the eigenvalues. This observation matches the theoretical results from Section 3.2.3 since we proved strong convergence in L^2 of the eigenfunctions and convergence of the eigenvalues to $\nu^{(m)}$. We also examine the eigenvalues and their ratios $\lambda_{1/L,h}^{(m)}/\lambda_{1/L,h}^{(m+1)}$ in Fig. 3.5, where the degeneracy of $m = 2$, pre-asymptotic effects, and a non-monotonic convergence is visible. This observation confirms the prediction of Corollary 3.1 that the fundamental ratio can only be uniformly bounded for all $L > L^*$ when pre-asymptotic effects have vanished.

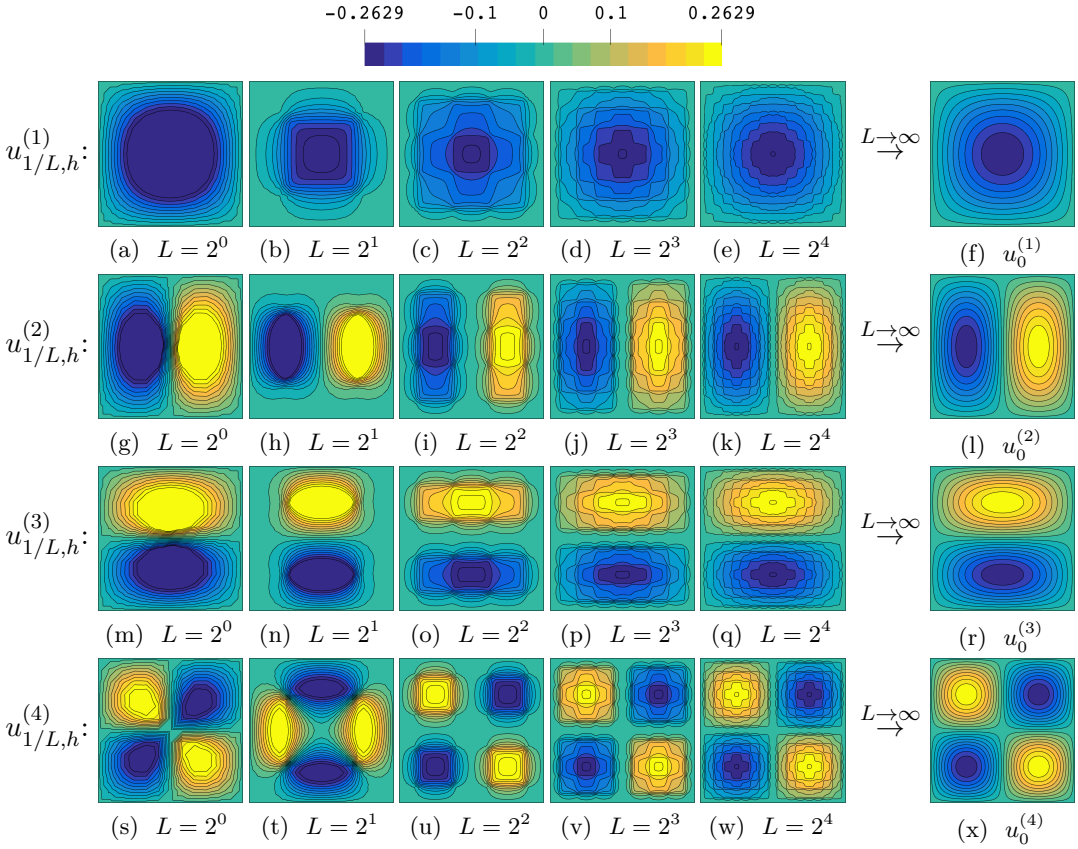


Figure 3.4: The first four calculated eigenfunctions of the eigenvalue homogenization problem (3.83) converge weakly for $L \to \infty$ to the solutions of the homogenized equation. The figure presents two-dimensional cut-planes through the middle of the domain at $y_1 = 1/2$.

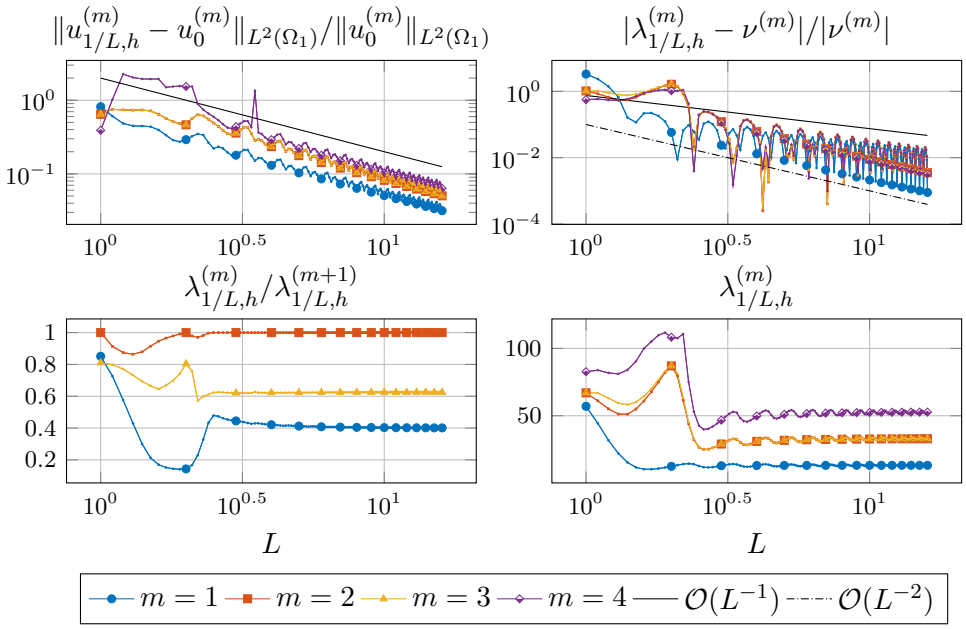


Figure 3.5: Errors between the solution of Eq. (3.83) and the corresponding homogenized limit: We can observe the first-order convergence for all eigenfunctions in the L^2-norm and at least first-order convergence for the eigenvalues. The ratios between two adjacent eigenvalues reveal a degenerate state and a non-monotonic convergence for the fundamental ratio $\lambda^{(1)}_{1/L,h}/\lambda^{(2)}_{1/L,h}$.

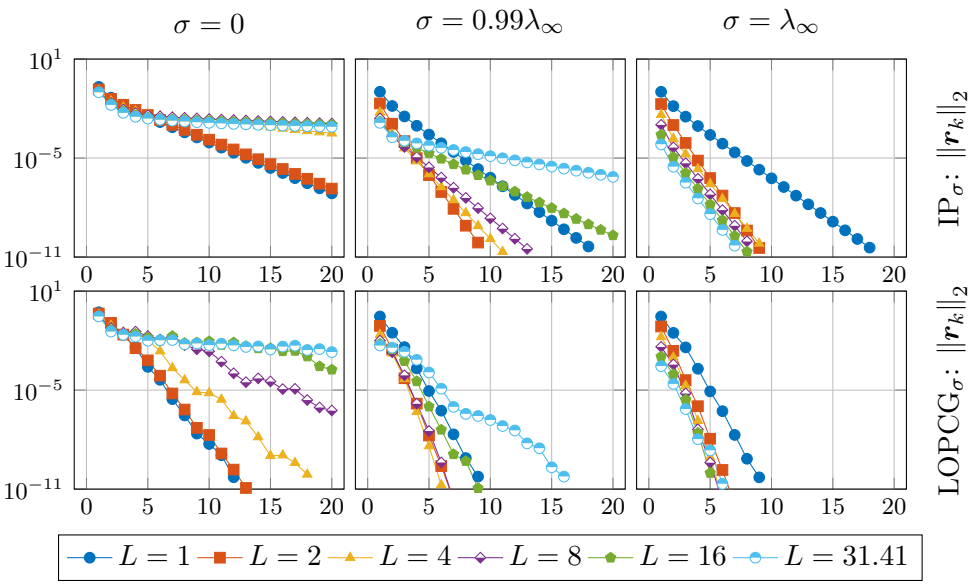
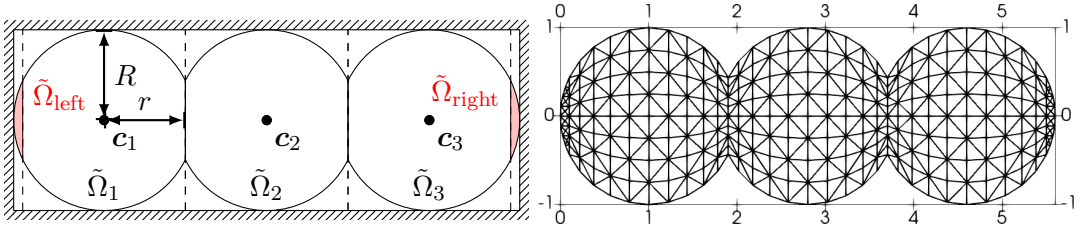


Figure 3.6: A comparison of the IP_σ and LOPCG_σ for the cases of $\sigma = 0$, $\sigma = 0.99\lambda_\infty$, and $\sigma = \lambda_\infty$ for different domain lengths L.

3.4.2 The Quasi-Optimal Shift-And-Invert Preconditioner

To show the practical advantage of using the quasi-optimal preconditioning technique of Section 3.3.2, we compare the convergence histories of the IP and LOPCG method for the cases of no shift ($\sigma = 0$), a good shift ($\sigma = 0.99\lambda_\infty$), and the quasi-optimal shift ($\sigma = \lambda_\infty$). We then aim to solve Eq. (3.2) on Ω_L for $\ell = 1$ and an increasing L. The quasi-optimal shift $\lambda_\infty = \lambda^{(1)}_{\mathcal{B}_\#,\mathcal{B}_d,1,V}(\Omega_1)$ is obtained in constant time for all L since it only depends on the fixed unit cell Ω_1. The calculations use $\mathbb{Q}_1$ finite elements on a regular mesh with mesh size $h = 1/100$ and the x-periodic potential $V(x, y) = 10^2 \sin(\pi x)^2 y^2$. We chose the start vectors $\boldsymbol{x}_0 = \mathbf{1}, \boldsymbol{x}_{-1} = \boldsymbol{e}_1$. The solvers aim to reduce the spectral residual $\boldsymbol{r}_k = \boldsymbol{A}\boldsymbol{x}_k - R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_k)\boldsymbol{B}\boldsymbol{x}_k$ below the tolerance $\texttt{TOL} = 10^{-10}$ and stop after 100 iterations. Both algorithms converged to the lowest eigenpair since the shifting strategy is order-preserving (c.f. Remark 3.4), and the start vector $\boldsymbol{x}_0$ can not be orthogonal to the non-sign-changing ground state.

The results in Fig. 3.6 indicate the drastic reduction in convergence speed for the unshifted algorithms. For the case of quasi-optimal preconditioning, both eigensolvers only need a couple of iterations to converge, as predicted by Theorem 3.3. When applying a good but not quasi-optimal shift of $0.99\lambda_\infty$, fast convergence rates for lower values of L can be observed. However, the convergence also deteriorates in the asymptotic limit of $L \to \infty$. This fact underlines the requirement for σ to be the exact asymptotic limit if the method shall provide convergence in a fixed number of iterations for all possible L. Furthermore, all three cases show a faster convergence of the LOPCG compared to the IP method.



(a) The geometry of $\tilde{\Omega}$. (b) A schematic discretization of $\tilde{\Omega}$.

Figure 3.7: A union of three disks ($R = 1$) domain with defects in the x-direction and overlap of $d = 0.1$: $\tilde{\Omega}$ comprises three identical unit cells $\tilde{\Omega}_i$ and two domain defects $\tilde{\Omega}_{\text{left}}, \tilde{\Omega}_{\text{right}}$.

3.4.3 Extension to Complex Domains: Barrier Principle and Defects in x-Direction

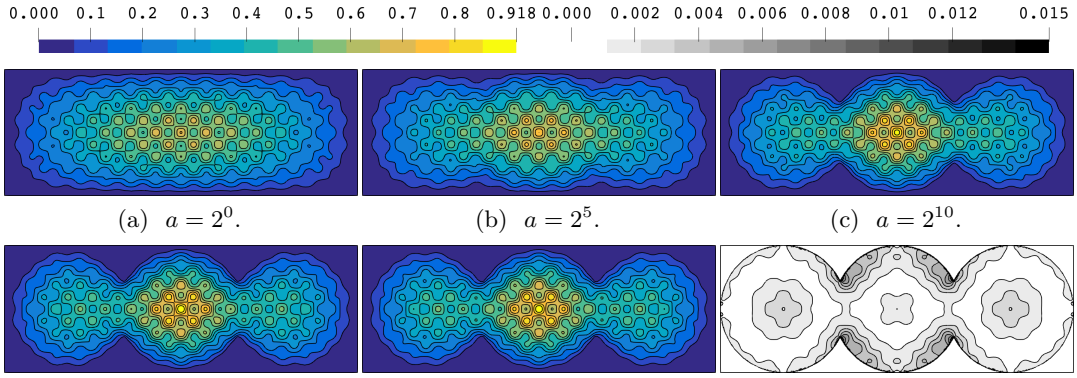
The initial box setup of $\Omega_L = (0, L)^p \times (0, \ell)^q$ from Section 3.1 is very suitable for the mathematical analysis performed in Section 3.2. However, we need to generalize the theory to more realistic domains for practical applications. Luckily, this can be quite intuitively done with some simple considerations.

Consider, for example, the setup of Fig. 3.7, in which one aims to simulate a union of three disks $\tilde{\Omega} = \bigcup_{i=1}^3 B_R((R + 2(i-1)r, 0)^T) = \tilde{\Omega}_{\text{left}} \cup \left(\bigcup_{i=1}^3 \tilde{\Omega}_i\right) \cup \tilde{\Omega}_{\text{right}}$ where $B_R(\boldsymbol{p})$ denotes a disk with radius R centered at $\boldsymbol{p}$ and $r = R - d$ with the overlap d. These disks are all aligned along the x-axis and have a fixed overlap. We define the rectangular unit cell as the box with side lengths $\{2r, 2R\}$, where one disk is contained entirely. Inside this unit cell, we assume the potential as directional periodic. Furthermore, we have domain defects $\tilde{\Omega}_{\text{left}}$ and $\tilde{\Omega}_{\text{right}}$ that are not part of any unit cell on the left and the right side. In this setup, two problems arise – the simulation of non-box-shaped domains and the handling of domain defects.

3.4.3.1 Barrier Principle for an Optical Lattice Potential

We could simulate the whole domain $\Omega_{L=6r+2d}$ to overcome the first issue. However, we are only interested in the union-of-disks domain $\tilde{\Omega}$, and a prescription of Dirichlet values on $\partial\tilde{\Omega}$ might be problematic since it is inside the domain. It is well known that we can simply modify the potential V to achieve this setting. To avoid nontrivial values of ϕ_L in certain regions, we can apply a significant penalty term to V. We call this strategy the *barrier principle*, which extends a given potential V to the barrier potential $\tilde{V}(\boldsymbol{z}; V, a) = V(\boldsymbol{z}) + a\chi_{\tilde{\Omega}^c}(\boldsymbol{z})$ where $a \geq 0$ is a penalty term. $\chi_{\tilde{\Omega}^c}$ is the indicator function for the complement of $\tilde{\Omega}$. For an increasing value of $a \to \infty$, we can still apply our theory for any finite value of a. In the limit case, the eigenvalue problem on the box-shaped domain Ω_{6r+2d} is equivalent to an eigenvalue problem, purely posed on the subdomain $\tilde{\Omega} \subset \Omega_{6r+2d}$.

To demonstrate the barrier effect of $\tilde{V}(\boldsymbol{z}; V, a)$, we inspect the union of three disks



(a) $a = 2^0$. (b) $a = 2^5$. (c) $a = 2^{10}$.

(d) $a = 2^{15}$. (e) $a = 2^{20}$. (f) $|(\phi|_{a=2^{25}} - \phi|_{a=2^{20}})|$.

Figure 3.8: Effect of the barrier potential $\tilde{V}(\boldsymbol{z}; 0, a)$ for varying penalty parameters a in the union-of-disks domain $\tilde{\Omega}$ of Fig. 3.7. With increasing a, the resulting problem statement reduces to the eigenproblem formulated in $\tilde{\Omega}$. When comparing the change between $a = 2^{15}$ and $a = 2^{20}$ in Fig. 3.8f, the solution's overall change is small and focused on the connection points. Also, we see an interpolation error at the disk boundary since the underlying mesh is not boundary-aligned.

case from Fig. 3.7 in combination with the optical lattice potential [146]

$$V(x, y) = 100 \left(1 - \sin \tfrac{\omega\pi(x-d)}{2(R-d)} \sin \tfrac{\omega\pi(y-(R-d))}{2(R-d)}\right), \tag{3.87}$$

where $\omega = 9, R = 1, d = 0.1$ for various penalty terms $a \in \{2^0, 2^5, \ldots, 2^{20}\}$. Fig. 3.8 shows the first eigenfunction, which is calculated with the LOPCG$_{\sigma=0}$ method for $\texttt{TOL} = 10^{-10}$ on a structured $\mathbb{Q}_1$-mesh with $h = 1/100$. It can be observed that with a increasing, the eigenfunction outside of $\tilde{\Omega}$ approaches zero. However, these above considerations are purely theoretical. In practice, we directly exclude the regions $\Omega_{6r+2d} \setminus \tilde{\Omega}$ and purely solve and mesh on $\tilde{\Omega}$ as displayed in Fig. 3.7b.

3.4.3.2 Principle of Defect Invariance

When it comes to simulating the domain of Section 3.4.3 with the quasi-optimally preconditioned eigensolver, we also need to calculate the shift $\sigma = \lambda_\infty$. For domains without any domain defect that perfectly match the potential's period, this is done by simulating one unit cell Ω_1 with Dirichlet zero in $\boldsymbol{y}$- and periodic boundary conditions in $\boldsymbol{x}$-direction as theoretically derived in Theorems 3.1 and 3.2.

For the case of defects located at the extremities of the expanding $\boldsymbol{x}$-direction as a subset of an imaginary unit cell Ω_1, we can do the same (if the potential acts as usual in the defect regions). The limit eigenvalue does not change since we can prove:

Theorem 3.4 (Principle of Defect Invariance)**.** *Let $\Omega_{L+2\delta}$ with L denoting the $\boldsymbol{x}$-period of the potential V and $\delta < L$ be given. For the linear Schrödinger eigenvalue problem Eq. (3.2) posed on $\Omega_{L+2\delta}$, it still holds that $\lim_{L\to\infty} \lambda_L^{(m)} = \lambda_{\varphi_y}$.*

Proof. Consider $\Omega_L \subset \Omega_{L+2\delta} \subset \Omega_{L+2}$. Then, by the inclusion principle [147, p13] for elliptic operators with Dirichlet boundary conditions, we have

$$\lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,0,V}(\Omega_L) \leq \lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,0,V}(\Omega_{L+2\delta}) \leq \lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,0,V}(\Omega_{L+2}) \tag{3.88}$$

Using the factorizations of Theorem 3.1, this is equivalent to

$$\lambda^{(1)}_{\varphi_y}(\Omega_L) + \lambda^{(m)}_{u_{y,2}}(\Omega_L) \leq \lambda^{(m)}_{\mathcal{B}_d,\mathcal{B}_d,0,V}(\Omega_{L+2\delta}) \leq \lambda^{(1)}_{\varphi_y}(\Omega_{L+2}) + \lambda^{(m)}_{u_{y,2}}(\Omega_{L+2}). \tag{3.89}$$

Since $\lambda^{(1)}_{\varphi_y}(\Omega_L) = \lambda^{(1)}_{\varphi_y}(\Omega_{L+2}) = \lambda^{(1)}_{\varphi_y}(\Omega_1)$ and $\lambda^{(m)}_{u_{y,2}}(\Omega_L), \lambda^{(m)}_{u_{y,2}}(\Omega_{L+2}) \in \mathcal{O}(1/L^2)$ by Theorem 3.2, we conclude with the Sandwich Lemma. □

We are now prepared to solve the union-of-disks geometry in the next Section 3.4.4.

3.4.4 Chain Model With Truncated Coulomb Potential in Two Dimensions

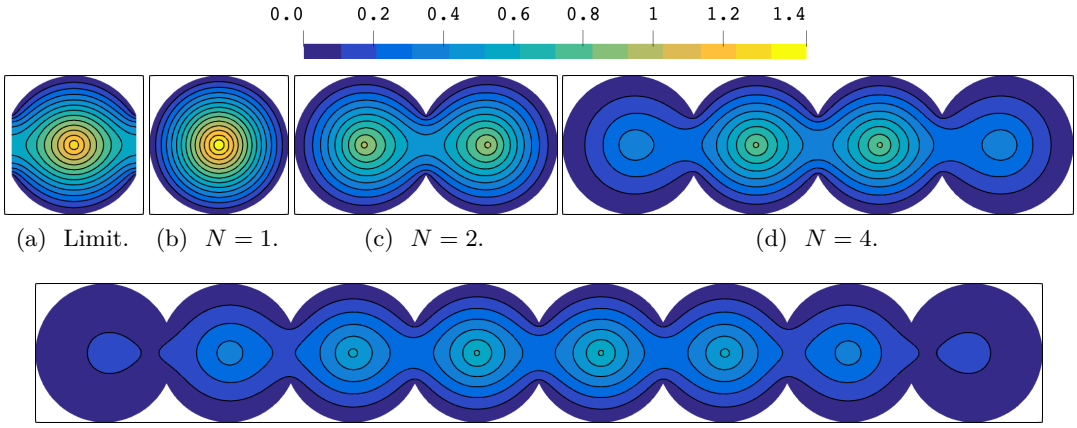
Consider the domain $\tilde{\Omega}_N = \bigcup_{i=1}^N B_R\left((R + 2(i-1)r, 0)^T\right)$ with the parameters $R = 1, r = 0.9$. Chain-like molecules in the context of molecular simulations inspire this model. For real applications, the potential is a Coulomb potential. Since a singularity of V violates the assumption (A2), we use a truncated Coulomb potential as

$$V_{\mathrm{C,lim}}(\boldsymbol{z}; b) = \begin{cases} -\frac{Z}{\|\boldsymbol{z}\|_2} & \text{for } \|\boldsymbol{z}\|_2 \geq b \\ -\frac{Z}{b} & \text{for } \|\boldsymbol{z}\|_2 < b \end{cases}, \tag{3.90}$$

to mimic, e.g., the electrostatic potential with charge $Z > 0$. We also have to neglect long-range interaction to fulfill the periodicity assumption on V. Consider the N centers $\{\boldsymbol{c}_i\}_{i=1}^N$ with $\boldsymbol{c}_i = (R + (i-1)2r, 0)$. We prescribe the compound periodic potential $V(\boldsymbol{z}) = \sum_{i \in \{i : |\boldsymbol{z} - \tilde{\boldsymbol{c}}_i| < R\}} V_{\mathrm{C,lim}}(\boldsymbol{z} - \tilde{\boldsymbol{c}}_i; b)$ where $\tilde{\boldsymbol{c}}_i \in \{\boldsymbol{c}_i\}_{i=1}^N \cup \{\boldsymbol{c}_1 - (2r, 0)\} \cup \{\boldsymbol{c}_N + (2r, 0)\}$ include ghost centers to fulfill the periodicity assumption (A1) also in the defect regions. We also note that the semi-positivity assumption (A3) is violated. However, it turned out that the resulting spectrum is still positive, the operator, thus, elliptic, and our theory is applicable. Nevertheless, it would also be possible to fulfill the assumption (A3) by adding a positive constant to the L^∞-potential without changing the resulting eigenfunctions.

A series of computations for $N = 1, 2, 4, \ldots, 32$ is performed using the LOPCG$_\sigma$ method for the potential parameters $Z = 1, b = 10^{-4}$ and the tolerance $\texttt{TOL} = 10^{-10}$. As shown in Fig. 3.7b, the spatial discretization uses unstructured but symmetrically meshed $\mathbb{P}_1$ elements. The quasi-optimal shift calculation uses the unit cell $\tilde{\Omega}_0 = \tilde{\Omega}_1 \cap [R - r, R + r] \times [-R, R]$ with periodic boundary conditions in x-direction and zero boundary conditions on the rest. Thus, $\sigma = \lambda_{\varphi_y}(\tilde{\Omega}_0) \approx 1.08784$ can be computed a priori in $\mathcal{O}(1)$ since the unit cell is independent of N.

In Fig. 3.9, the resulting ground state eigenfunctions $\phi_h^{(1)}$ are presented with the solution to the base problem in Fig. 3.9a. We can observe for increasing N that the solution inside a single disk approaches the shape of the solution to the unit



(a) Limit. (b) $N = 1$. (c) $N = 2$. (d) $N = 4$.

(e) $N = 8$.

Figure 3.9: Contours of the first eigenfunction for the union of N disks using the truncated Coulomb potential without long-range interactions: Fig. 3.9a shows the asymptotic limit eigenfunction with periodic boundary conditions in the x-direction.

cell problem. This convergence is the expected behavior, as shown theoretically in Section 3.2. In Table 3.1, we observe that, due to the quasi-optimal preconditioning, the number of iterations needed to meet the required residual tolerance is of order $\mathcal{O}(1)$. Furthermore, the method shows a linearly scaling behavior since the ratio of calculation time to the disk amount t_{eig}/N seems to be independent of N.

3.4.5 Plane Model With Kronig–Penney Potential in Three Dimensions

Finally, we also show the preconditioner's quasi-optimality for a three-dimensional case with two expanding directions ($p = 2$, $q = 1$). We use a three-dimensional Kronig–Penney potential [258], defined by $V(\boldsymbol{z}) = 0$ for $\|\boldsymbol{z} \bmod 1 - \mathbf{1}/2\|_1 < 1/4$ and $V(\boldsymbol{z}) = 100$ otherwise, where $\mathbf{1}$ denotes the vector of ones in d dimensions. This potential represents cubic wells with sidelength 0.5 centered in the unit cubes, which form the plane-like expanding domain $\Omega_L = (0, L)^2 \times (0, 1)$ with $L \in \mathbb{N}$. We again calculate the quasi-optimal shift $\sigma = \lambda^{(1)}_{\mathcal{B}_\#, \mathcal{B}_d, 1, V}(\Omega_1)$ on the unit cube and use it to precondition the Ω_L-problem. Finally, both problems are discretized using a uniform mesh size of $h = 1/10$ and $\mathbb{Q}_1$ elements resulting in $\sigma \approx 57.60485$. We use the LOPCG$_\sigma$ method with $\texttt{TOL} = 10^{-10}$ and solve for the ground state solution. The simulations are performed on a series of domains Ω_L with $L \in \{1, 2, 4, \ldots, 32\}$. In Table 3.2, we observe that the number of eigensolver iterations k_{it} does not increase for $L \to \infty$, confirming our theory. However, in contrast to Table 3.1, a slight increase in solution time per number of unit cells (t_{eig}/L^2) can be observed. This increase is the expected behavior of using a direct solver for sparse matrices with increased bandwidth for $L \to \infty$ for our case of $p = 2$ expanding directions.

Table 3.1: The summary of computations for the union of N disks with the truncated Coulomb potential. Due to the same discretization density, the number of nodes n_{nodes} for each mesh is approximately proportional to the number of disks N (up to the defects). The wall times are measured on an Intel X7542 CPU using one core.

$N \propto L$	n_{nodes}	$\lambda_h^{(1)}$	$\max \phi_h^{(1)}$	k_{it}	$\frac{\lambda_{\max}(\boldsymbol{A}-\sigma\boldsymbol{B})}{\lambda_{\min}(\boldsymbol{A}-\sigma\boldsymbol{B})}$	t_{eig} $[s]$	$\frac{t_{\text{eig}}}{N}$ $[s]$
1	89,869	1.96222	1.44	5	$4.83 \cdot 10^5$	1.83	1.83
2	$1.69 \cdot 10^5$	1.46912	0.96	5	$1.25 \cdot 10^6$	3.95	1.97
4	$3.27 \cdot 10^5$	1.20013	0.85	5	$4.25 \cdot 10^6$	7.7	1.92
8	$6.44 \cdot 10^5$	1.11768	0.64	5	$1.14 \cdot 10^7$	17.03	2.13
16	$1.28 \cdot 10^6$	1.0955	0.46	5	$2.55 \cdot 10^7$	33.06	2.07
32	$2.54 \cdot 10^6$	1.08978	0.33	5	$5.33 \cdot 10^7$	64.96	2.03

Table 3.2: The summary of computations for the plane-like expanding domain in three directions with the Kronig–Penney potential. The number of unit cells N now scales quadratically with L.

L	n_{nodes}	$\lambda_h^{(1)}$	$\max \phi_h^{(1)}$	k_{it}	t_{eig} $[s]$	$\frac{t_{\text{eig}}}{L^2}$ $[s]$
1	1,331	58.99915	4.71	5	0.16	0.16
2	4,851	58.30881	2.31	6	0.52	0.13
4	18,491	57.81186	1.95	6	3.01	0.19
8	72,171	57.6587	1.09	7	15.28	0.24
16	$2.85 \cdot 10^5$	57.61845	0.56	7	73.33	0.29
32	$1.13 \cdot 10^6$	57.60826	0.28	6	320.64	0.31

3.5 Conclusion

This chapter presented a quasi-optimal shift-and-invert preconditioner to solve the linear periodic Schrödinger eigenvalue problem in a constant number of eigensolver iterations for domains expanding periodically in a subset of directions. First, we analyzed and proved the quasi-optimality of the method using factorization and homogenization techniques. The analysis revealed powerful insights into the behavior of the eigenfunctions and eigenvalues. Significantly, the representation of the searched eigenfunction as the product of easy-to-calculate functions leads to a decisive result – the corresponding eigenvalues can be expressed as the sum of other eigenvalues, which can be much easier computed in practice than solving the whole system. This realization makes the proposed method very practical since calculating the quasi-optimal shift can be done in $\mathcal{O}(1)$. We then extended the results to complex and defect domain shapes to allow for a broader range of geometrical applications. Finally, in our experiments, we showed the practical usability of the method for chain-like and plane-like expanding domains.

Limitations of our method include the assumptions on the potential V to be

essentially bounded and periodic. Also, we observed that using the perfect shift in the eigensolver algorithms leads naturally to an ill-conditioned system matrix $(\boldsymbol{A} - \sigma \boldsymbol{B})$. Thus, future work could weaken the periodicity assumptions on V by, e.g., allowing for a perturbation of δV that vanishes in the limit $L \to \infty$. Also, efficient solvers for the linear system involving $(\boldsymbol{A} - \sigma \boldsymbol{B})$ must be constructed when the system size requires iterative linear solvers.

4 PerFact: A Scalable Two-Level Domain Decomposition Eigensolver for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains

Accelerating iterative eigenvalue algorithms is often achieved by employing a spectral shifting strategy. Unfortunately, improved shifting typically leads to a smaller eigenvalue for the resulting shifted operator, which in turn results in a high condition number of the underlying solution matrix, posing a major challenge for iterative linear solvers. This chapter introduces a two-level domain decomposition preconditioner that addresses this issue for the linear Schrödinger eigenvalue problem, even in the presence of a vanishing eigenvalue gap in non-uniform, expanding domains. Since the quasi-optimal shift, which is already available as the solution to a spectral cell problem, is required for the eigenvalue solver, it is logical to also use its associated eigenfunction as a generator to construct a coarse space. We analyze the resulting two-level additive Schwarz preconditioner and obtain a condition number bound that is independent of the domain's anisotropy, despite the need for only one basis function per subdomain for the coarse solver. Several numerical examples are presented to illustrate its flexibility and efficiency.

—

This chapter has been submitted and published as a preprint [247]:

- L. Theisen and B. Stamm. *A Scalable Two-Level Domain Decomposition Eigensolver for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains*. Submitted. 2023. DOI: 10.48550/arXiv.2311.08757. arXiv: 2311.08757 [cs, math]

4.1 Introduction

In this chapter, we present a robust and efficient method to solve parametrized linear Schrödinger eigenvalue problems (EVPs) on open bounded domains $\Omega_L \subset \mathbb{R}^d$ of the form: Find $(\phi_L, \lambda_L) \in (H_0^1(\Omega_L) \setminus \{0\}) \times \mathbb{R}$, such that

$$-\Delta\phi_L + V\phi_L = \lambda_L\phi_L \quad \text{in } \Omega_L, \tag{4.1}$$

where $H_0^1(\Omega_L)$ denotes the standard Sobolev space of index 1 with zero Dirichlet trace on $\partial\Omega_L$. The main focus will be on anisotropically expanding domains Ω_L, which are modeled by d-dimensional boxes given by

$$\boldsymbol{z} \in \Omega_L = (0, L)^p \times (0, \ell)^q =: \Omega_{\boldsymbol{x}} \times \Omega_{\boldsymbol{y}} \subset \mathbb{R}^d \quad \text{with } L, \ell \in \mathbb{R}, \tag{4.2}$$

where the spatial variables are collected as $\boldsymbol{z} := (\boldsymbol{x}, \boldsymbol{y}) = (x_1, \ldots, x_p, y_1, \ldots, y_q)$, highlighting the fact that some directions expand with $L \to \infty$ while the other directions are fixed with $\ell = \text{const}$. Note that the box setup is only chosen for simplicity of the analysis, and we provide an elegant extension in Section 4.5 to the general case. For the external potential V in Eq. (4.1), we assume the following:

(B1) The potential V is directional-periodic with a period of 1 in each expanding direction: $V(\boldsymbol{x}, \boldsymbol{y}) = V(\boldsymbol{x} + \boldsymbol{i}, \boldsymbol{y})$ for all $(\boldsymbol{x}, \boldsymbol{y}) \in \Omega_L, \boldsymbol{i} \in \mathbb{Z}^p$;

(B2) The potential V is essentially bounded: $V \in L^\infty(\Omega_L)$.

The period of 1 in (B1) is only chosen for simplicity and (B2) allows us to assume $V \geq 0$ a.e. in Ω_L since a constant spectral shift to V does not affect the eigenfunctions of Eq. (4.1). Fig. 4.1a presents the geometrical framework and indicates the properties of V. After discretization, the problem (4.1) leads to a large, sparse, generalized algebraic EVP of the form

$$\boldsymbol{A}\boldsymbol{x} = \lambda \boldsymbol{B}\boldsymbol{x}, \tag{4.3}$$

that is then solved with classical iterative algorithms like, for example, the inverse power method (IPM) [243] or the Locally Optimal Preconditioned Conjugate Gradient (LOPCG) [167] method. However, the convergence speed of these algorithms depends on the *fundamental ratio* $r_L := \lambda_L^{(1)}/\lambda_L^{(2)}$. And for $L \to \infty$, the *fundamental gap* collapses (i.e., $\lambda_L^{(1)} - \lambda_L^{(2)} \to 0$), leading to $r_L \to 1$ as we extensively illustrate in the motivational Section 4.1.2. As a result, the convergence speed of iterative eigensolvers, equal to the fundamental ratio, deteriorates drastically and becomes arbitrarily bad.

This problem was overcome in Chapter 3 by the quasi-optimal shift-and-invert (QOSI) preconditioner that replaces the original matrix from Eq. (4.3) by a shifted $\boldsymbol{A}_\sigma := \boldsymbol{A} - \sigma\boldsymbol{B}$. The shifting parameter $\sigma \in \mathbb{R}$ is set to be the L-asymptotic limit of the first eigenvalue, i.e. $\sigma := \lim_{L\to\infty} \lambda_L^{(1)}$, which can be easily computed by solving the same operator on a unit cell domain Ω_1 where periodic conditions replace all Dirichlet boundary conditions in the expanding $\boldsymbol{x}$-direction as it was shown in Chapter 3. More precisely, (ψ, σ) is the first eigenpair to the problem

$$\begin{cases} -\Delta\psi + V\psi = \sigma\psi \text{ in } \Omega_1 = (0,1)^p \times (0,\ell)^q, \\ \qquad\qquad \psi = 0 \text{ on } \Omega_{\boldsymbol{x}} \times \{0, \ell\}^q, \\ \qquad\qquad \psi \text{ and } \partial_{x_i}\psi \text{ are } x_i\text{-periodic.} \end{cases} \tag{4.4}$$

The QOSI preconditioner $\boldsymbol{A}_\sigma^{-1}$ then leads to a uniformly bounded shifted fundamental ratio as, for some $L^* \in \mathbb{R}$, $r_L(\sigma) := (\lambda_L^{(1)} - \sigma)/(\lambda_L^{(2)} - \sigma) \leq C < 1$ for all $L > L^*$ such

that the eigensolver converges in $k_{\text{it}} \in \mathcal{O}(1)$ iterations (i.e., up to a multiplicative constant optimal).

However, in each eigensolver iteration, c.f. inverse iteration $\boldsymbol{x}_{k+1} = \boldsymbol{A}_\sigma^{-1}\boldsymbol{B}\boldsymbol{x}_k$, the application of $\boldsymbol{A}_\sigma^{-1}$ amounts to solving shifted linear systems of the form

$$(\boldsymbol{A} - \sigma\boldsymbol{B})\boldsymbol{x} = \boldsymbol{b}. \tag{4.5}$$

The catch here is that in the critical limit, $L \to \infty$, with fixed discretization $h = \text{const}$, $\lambda_L^{(1)} \to \sigma$ by construction and thus the condition number $\kappa(\boldsymbol{A} - \sigma\boldsymbol{B}) = (\lambda_L^{(n)} - \sigma)/(\lambda_L^{(1)} - \sigma) \to \infty$ as observed in Chapter 3. The issue of an exploding condition number is dramatic when we employ iterative linear solvers like the conjugate gradient (CG) [151] method to solve Eq. (4.5) since their convergence rate has now become arbitrarily bad.

4.1.1 Our Contribution

This chapter fixes these problems by introducing and analyzing a new preconditioner $\boldsymbol{M}_\sigma^{-1}$, now for the shifted linear systems of Eq. (4.5). As it is very natural for the present geometry Ω_L and potentials V, the preconditioner uses classical, overlapping Schwarz domain decomposition (DD) [98]. However, a coarse space must be prescribed for robustness and numerical scalability. Surprisingly, it is even directly available since it is based on the spectral asymptotics of the problem. Since the coarse space components are inherently related to the function ψ of Eq. (4.4) used in the so-called *factorization principle* (c.f. [8] and Section 3.2.2) to derive σ, we call $\boldsymbol{M}_\sigma^{-1}$ the periodic factorization (PerFact) preconditioner. The PerFact-preconditioned shifted linear systems can then also be solved in quasi-optimal iterations since $\kappa(\boldsymbol{M}_\sigma^{-1}\boldsymbol{A}_\sigma) \leq C$ for all L. As a result, only the unique combination of the QOSI eigen-preconditioner and the PerFact linear-preconditioner, in the end, yields an efficient iterative algorithm that is robust for all domain sizes L.

For the provided analysis, we embed the preconditioner partly in the theory of spectral coarse spaces while still focussing on the unusual case of anisotropically expanding domains. This setup, and the challenging fact that we asymptotically shift with the lowest eigenvalue, required a new SPSD splitting and the introduction of an auxiliary periodic neighborhood decomposition.

4.1.2 Motivation: The Shifting Dilemma in the Laplace EVP

When it comes down to highlighting the critical difficulty and motivating the present work, the Laplace eigenvalue problem with $V(\boldsymbol{z}) = 0$ in Eq. (4.1) is the perfect academic example fulfilling the assumptions (B1) and (B2). For a standard second-order five-point finite difference stencil of a two-dimensional Ω_L with $p, q = 1$ while $\ell = 1$, the eigenvalues of the resulting discretization matrix $\boldsymbol{A}$ are all known. Let n denote the number of grid points per unit length and introduce the uniform mesh size $h := 1/(n+1)$. Then all eigenvalues are given (c.f. the formula in [89, p40, p63]

using $2\sin^2(x/2) = 1 - \cos x$) by

$$\lambda_{i,j} \in \left\{ \tfrac{4}{h^2} \left(\sin^2(\tfrac{\pi h i}{2L}) + \sin^2(\tfrac{\pi h j}{2}) \right) \middle| \; i \in \{1, \dots, L(n+1)-1\}, j \in \{1, \dots, n\} \right\}. \quad (4.6)$$

If we now employ the σ-shifted inverse power method ($\boldsymbol{x}_{k+1} := \mathcal{R}(\boldsymbol{A}_\sigma^{-1}\boldsymbol{x}_k)$ with some retraction $\mathcal{R} : \mathbb{R}^n \mapsto \mathbb{S}^{n-1}$ onto the unit sphere) for $\boldsymbol{A}_\sigma^{-1} = (\boldsymbol{A} - \sigma\boldsymbol{I})^{-1}$ as the outer iteration, the convergence rate is given by the fundamental ratio [129, p366] of the first two eigenvalues $\rho_{\text{IPM}} = r_L(\sigma) = \frac{\lambda_1 - \sigma}{\lambda_2 - \sigma}$ where (for $L > \ell = 1$)

$$\lambda_1 := \lambda_{1,1} = \tfrac{4}{h^2} \left(\sin^2(\tfrac{\pi h}{2L}) + \sin^2(\tfrac{\pi h}{2}) \right), \; \lambda_2 := \lambda_{2,1} = \tfrac{4}{h^2} \left(\sin^2(\tfrac{2\pi h}{2L}) + \sin^2(\tfrac{\pi h}{2}) \right). \quad (4.7)$$

On the other hand, in each outer iteration, a large, sparse, linear system with $\boldsymbol{A}_\sigma$ needs to be solved. Using the CG method, the convergence rate of that inner solver is given by $\rho_{\text{CG}} = (\sqrt{\kappa(\boldsymbol{A}_\sigma)} - 1)/(\sqrt{\kappa(\boldsymbol{A}_\sigma)} + 1)$ with

$$\kappa(\boldsymbol{A}_\sigma) = (\lambda_{\max} - \sigma)/(\lambda_1 - \sigma), \text{ and } \lambda_{\max} = \lambda_{L(n+1)-1,n} = \tfrac{4}{h^2} \left(\cos^2(\tfrac{\pi h}{2L}) + \cos^2(\tfrac{\pi h}{2}) \right). \quad (4.8)$$

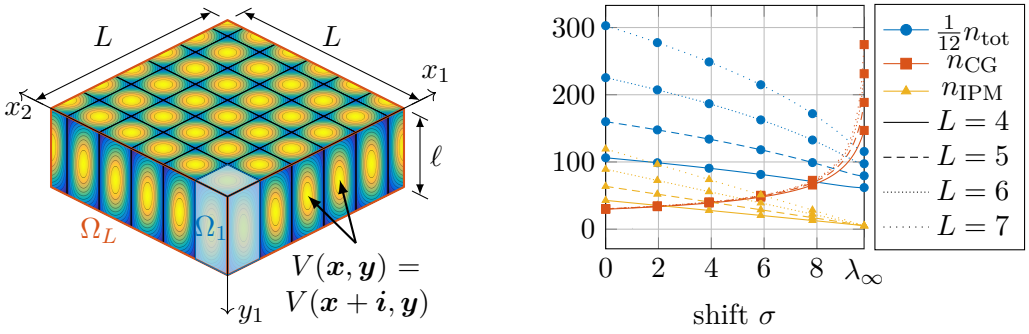
Now, the error for the first eigenvector in the IPM reduces [129, Thm. 8.2.1] as $|\sin\theta_k| \leq \rho_{\text{IPM}}^k \tan\theta_0$ with $\theta_k \in [0, \frac{\pi}{2}]$ being the angle of the current iterate $\boldsymbol{x}_k$ to the real eigenvector $\boldsymbol{x}^{(1)}$ defined by $\cos(\theta_k) = |\langle \boldsymbol{x}^{(1)}, \boldsymbol{x}_k \rangle|$. So, in order to decrease the ratio $|\sin\theta_k|/\tan\theta_0$ by a factor of $1/R \in \mathbb{R}$, we need $n_{\text{IPM}} = -\ln R / \ln \rho_{\text{IPM}}$ iterations where $\ln(\cdot)$ denotes the natural logarithm. Each application of $\boldsymbol{A}_\sigma$ is then executed using the CG algorithm until the residual reduction $\|\boldsymbol{r}_k\|_2 / \|\boldsymbol{r}_0\|_2 \leq 1/Q \in \mathbb{R}$ is archived, where $\boldsymbol{r}_k$ is the residual in step k. Using the residual-to-error-energy bound [162, p4] and [253, Lem. C.10], we obtain the relation

$$\frac{\|\boldsymbol{r}_k\|_2}{\|\boldsymbol{r}_0\|_2} \leq \sqrt{\kappa(\boldsymbol{A}_\sigma)} \frac{\|\boldsymbol{e}_k\|_{\boldsymbol{A}_\sigma}}{\|\boldsymbol{e}_0\|_{\boldsymbol{A}_\sigma}} \leq 2\sqrt{\kappa(\boldsymbol{A}_\sigma)} \rho_{\text{CG}}^k, \quad (4.9)$$

in which $\boldsymbol{e}_k$ is the error between the current inner iterate and the linear system's solution. Thus, each inner CG algorithm needs $n_{\text{CG}} = -\ln(2\sqrt{\kappa(\boldsymbol{A}_\sigma)}Q)/\ln\rho_{\text{CG}}$ iterations. Assuming no prior information and equally good initial guesses, the inner-outer algorithm, then, in total, needs $n_{\text{tot}} = n_{\text{IPM}} n_{\text{CG}}$ CG iterations, which is a good measure of the computational cost.

For an algorithm to be efficient, it must be scalable w.r.t. L. Thus, doubling the system size (i.e., $L \mapsto 2L$) while doubling computational resources (i.e., ranks or processes) leads ideally to the same wall clock time needed for the algorithm. Assuming perfect parallelizability of all operations and because the system matrix is twice as large in that case, the wall clock time can only be kept constant if the total number of operations does not increase when doubling L. However, even in this simple example, we observe the *shifting dilemma* for increasing domain lengths $L \to \infty$ and constant h since we obtain the following for the different possible shifts:

- **No shift** ($\sigma = 0$): The application of no shift results in the fundamental ratio converging to 1 since $\lim_{L\to\infty} \lambda_1/\lambda_2 = 1$ using the values from Eq. (4.7).



(a) Geometric setup of Ω_L and V.

(b) Iteration numbers for different shifts σ.

Figure 4.1: **(a)** Geometric setup of Ω_L with $p = 2$ expanding and $q = 1$ fixed directions with dimensions $L = 5.5, \ell = 2$. **(b)** Iteration number estimates for an inner-outer eigenvalue algorithm using IPM/CG for the Laplacian EVP on $(0, L) \times (0, 1)$ using finite differences ($h = 1/10$) and different shifts σ. Note that the arbitrary scaling of n_{tot} is only applied for better visualization.

The condition number $\kappa(\boldsymbol{A})$, on the other hand, converges to a finite value as $\lim_{L\to\infty} \lambda_{\max}/\lambda_1 = (1 + \cos^2(\frac{\pi h}{2}))/\sin^2(\frac{\pi h}{2})$ using Eqs. (4.7) and (4.8). Thus, for $L \to \infty$, we obtain $n_{\text{IPM}} \to \infty$ and $n_{\text{CG}} \to C > 0$, and the algorithm can not be scalable as $n_{\text{tot}} \to \infty$.

- **QOSI shift** from Chapter 3 ($\sigma = \lambda_\infty = \frac{4}{h^2}\sin^2(\frac{\pi h}{2})$): Using the quasi-optimal shift leads to a uniform bounded ratio $\frac{\lambda_1 - \sigma}{\lambda_2 - \sigma} = \sin^2(\frac{\pi h}{2L})/\sin^2(\frac{2\pi h}{2L}) \to 1/4$ for $L \to \infty$ using l'Hôpital's rule and thus bounded outer itations with $n_{\text{IPM}} \leq C$. However, the condition number $\kappa(\boldsymbol{A}_\sigma)$ of the shifted matrix now explodes when $L \to \infty$ as $\lim_{L\to\infty}(\lambda_{\max} - \sigma)/(\lambda_1 - \sigma) > \lim_{L\to\infty}(\cos^2(\frac{\pi h}{2L})/\sin^2(\frac{\pi h}{2L})) = \infty$ by the minorant argument. This behavior is natural since we shifted by the asymptotic limit of the first eigenvalue λ_1. Thus, for $L \to \infty$, $n_{\text{CG}} \to \infty$, and $n_{\text{tot}} \to \infty$.

- **Adaptive hybrid shift** ($\sigma \mapsto \sigma_L \in (0, \lambda_\infty)$): If both extreme cases do not provide a scalable method, one might ask if there is a chance that a (possible L-dependent) shift in the whole interval $(0, \lambda_\infty)$ results in n_{tot} to be bounded w.r.t. L. In Fig. 4.1b, we evaluated the formulas for n_{IPM} and n_{CG} using an exemplary setup of $h = 1/10, R = e^7, Q = e^4$ to obtain values for n_{tot} depending on the shift σ for different $L \in \{4, 5, 6, 7\}$. We observe that for all σ, the total CG iterations n_{tot} grow as $L \to \infty$. Thus, there is also no sweet spot in between both extremal cases, which suggest that also no adaptive shift can produce a scaling algorithm since all possible shifts still result in $n_{\text{tot}} \to \infty$.

Thus, the shifting dilemma prevents us from using the QOSI strategy with standard inner solvers without special treatment, and we need to provide something better –

another preconditioner for the inner CG solver. In our discussion, we intentionally omitted the consideration of the case where $h \to 0$. Including this scenario would obscure the primary insight that shifting is why we require an alternative preconditioner in this context, in contrast to the classical case of a fixed domain and finer meshes.

4.1.3 State-of-the-art and Context

We can integrate our results into the existing corpus of related research in four aspects – the anisotropic geometry, the studied model equation, the usage of domain decomposition algorithms, and the construction of spectral coarse spaces.

The Schrödinger equation Eq. (4.1) describes the stationary states of the wave function ϕ within a quantum mechanical system under the influence of the external potential V. Although this linear equation is only suitable for solving simple systems, it is also essential for, e.g., more complex simulations in computational chemistry. In fact, for nonlinear eigenvalue problems, the celebrated method of self-consistent field (SCF) iterations [60, 226], e.g., relies on linear eigenvalue solvers for equations of the present type – even in each step of the nonlinear iteration loop. One applicable example is the Gross-Pitaevski equation (GPE) for modeling Bose–Einstein condensates, where one is typically interested in the ground state with minimal energy. In contrast to SCF-like schemes, direct minimization aims to minimize the associated energy to obtain the ground-state solution of the system directly. Methods of this category are, e.g., based on gradient flow [102, 139, 146], manifold optimization [15, 19, 64], preconditioned CG [117], or other variations [18, 143], and all have one thing in common: they repeatedly need to solve systems similar to Eq. (4.5) – either for the update step directly or implicitly when applying a preconditioner.

Second, although the considered anisotropic and periodic geometrical setup seems somewhat artificial initially, it is precisely this geometry where very exceptional phenomena happen. One example is the carbon allotope hierarchy, where the dimensions successively increased from the *0d* fullerene [177] to *1d* carbon nanowires [271] or -tubes [34, 157], resulting in the Nobel Prize-winning works on *2d* graphene [126, 213, 214]. More recent developments now consider *2.5d* materials by combining two or more periodic but twisted material sheets to create Moiré superlattices [62] such as twisted bilayer graphene [66, 67] or increasing the dimension above $d > 3$ in the framework of time crystals [175, 265] ($p = 3, q = 1$ in Eq. (4.2)). Understanding the properties of crystalline structures with periodic operators [63, 150] and the efficient search for promising structures, e.g., based on high-throughput simulations [149, 201], are therefore enormously relevant.

Third, the considered geometries of Eq. (4.2) naturally harmonize very well with the DD method. In fact, for elliptic source problems, the classical Schwarz method is weakly scalable and does not require coarse correction [65] on anisotropic domains with Dirichlet boundaries. This remarkable fact was studied in [86, 87, 88] and further extended in [90, 91, 136, 224], directly inspiring this work with the simple question: What happens to the DD algorithms when we replace the linear problem with an eigenvalue problem for the same operator? *It changes a lot* – is the short answer,

which we elaborate on in this chapter. Compared to multigrid methods, the DD approach benefits our setup since very efficient local solvers are available (c.f. [193]). Using DD for eigenvalue computations has been introduced previously. There are approaches based on local energy minimization with a global coupling strategy, such as the Multilevel Domain Decomposition (MDD) [39, 40], the Automated Multilevel Substructuring (AMLS) method [42], Divide-and-Conquer (DAC) approaches [74], Schwarz-type sequential optimization methods [196, 197, 198], or other related work [97, 159, 269]. We, however, follow the inner-outer paradigm since we already have access to a quasi-optimal eigensolver from Chapter 3 and use the DD strategy to solve the arising shifted linear systems. This strategy allows us to show off an inherent connection between the preconditioning of eigenvalue solvers and the construction of coarse spaces for linear solvers, which, to our knowledge, has yet to be discussed in the literature so far.

Coarse spaces for overlapping Schwarz methods are, thus, the last to be mentioned. Spectral coarse spaces, in particular, usually deal with high-contrast diffusion problems by using local cell problems in the volume [36, 119, 120], in the overlap [235, 236], or based on DtN maps [99] to construct efficient coarse basis functions based on spectral problems. Recent developments try to extend these approaches to the algebraic case [2, 3, 130] and establish the connection to multiscale methods in general [85, 140, 141]. Thus, multiscale methods are naturally related and have been used successively for the GPE [144, 145, 220].

4.1.4 Outline of the Chapter

In the following Section 4.2, we present our two-level algorithm after introducing inner-outer eigenvalue algorithms in Section 4.2.1, their usage with two-level domain decomposition methods in Section 4.2.2, and the definition of the new coarse space in Section 4.2.3. After illustrating the connection to spectral coarse spaces, we analyze the coarse space and provide a condition number bound in Section 4.3. Finally, Section 4.5 presents various numerical examples to test the method with full generality.

4.2 Domain Decomposition for Eigenvalue Algorithms

For the numerical solution of the Schrödinger equation (4.1), we first apply a classical Galerkin finite element scheme. Let $\mathcal{T}_h$ be a conforming and shape-regular partition of the domain Ω_L into finite elements $\tau \in \mathcal{T}_h$, where $h := \max_{\tau\in\mathcal{T}_h} \operatorname{diam}\tau$ [50, p150]. We then define the finite element subspace $V_h(\Omega_L) \subset H_0^1(\Omega_L), |V_h(\Omega_L)| = n$, that consists of piecewise polynomials with total degree r from the space of polynomials $\mathcal{P}_r$ as $V_h(\Omega_L) := \{u \in H_0^1(\Omega_L) \,|\, u|_\tau \in \mathcal{P}_r \,\forall \tau \in \mathcal{T}_h\}$. The resulting discrete problem then searches for the function $\phi_h \in V_h(\Omega_L) \setminus \{0\}$ and the value $\lambda_h \in \mathbb{R}$ such that

$$\forall v_h \in V_h(\Omega_L): \quad \int_{\Omega_L} \nabla\phi_h \cdot \nabla v_h \,\mathrm{d}\boldsymbol{z} + \int_{\Omega_L} V\phi_h v_h \,\mathrm{d}\boldsymbol{z} = \lambda_h \int_{\Omega_L} \phi_h v_h \,\mathrm{d}\boldsymbol{z}. \tag{4.10}$$

For the finite element basis of $V_h(\Omega_L)$, denoted by $\{N_i\}_{i=1}^n$, we have $\phi_h = \sum_{i=1}^n x_i N_i$ where we collect all coefficients in the vector $\boldsymbol{x}_h \in \mathbb{R}^n$. Thus, Eq. (4.10) is equivalent to the algebraic generalized eigenvalue problem

$$\boldsymbol{A}\boldsymbol{x}_h = \lambda_h \boldsymbol{B}\boldsymbol{x}_h, \tag{4.11}$$

in which $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is the sum of the usual Laplacian stiffness matrix and the contribution from the potential V, while $\boldsymbol{B} \in \mathbb{R}^{n\times n}$ is the usual L^2-mass matrix. In order to solve the GEVP (4.11) for the lowest eigenpair $(\boldsymbol{x}_h^{(1)}, \lambda_h^{(1)})$, we can employ an inverse power method. This algorithm, starting from an initial guess of $\boldsymbol{x}_0 \in \mathbb{R}^n$, successively applies the inverse of the operator as $\boldsymbol{x}_{k+1} = \boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{x}_k$ and $\boldsymbol{B}$-normalizes the result. As motivated in Section 4.1.2 for the Laplcian EVP and proved in Chapter 3 for the Schrödinger EVP with the assumptions (B1) and (B2), we need to apply shifting to obtain robustness for the domain size L. This preconditioning is archived by calculating the quasi-optimal shift σ as the first eigenvalue of the *limit cell problem* (4.4) posed on the unit cell $\Omega_1 \subset \Omega_L$: Find $(\psi_h, \sigma) \in (V_{h,\#}(\Omega_1) \setminus \{0\}) \times \mathbb{R}$ such that

$$\forall v_h \in V_{h,\#}(\Omega_1): \quad \int_{\Omega_1} \nabla\psi_h \cdot \nabla v_h \, \mathrm{d}\boldsymbol{z} + \int_{\Omega_1} V\psi_h v_h \, \mathrm{d}\boldsymbol{z} = \sigma \int_{\Omega_1} \psi_h v_h \, \mathrm{d}\boldsymbol{z}, \tag{4.12}$$

in which $V_{h,\#}(\Omega_1) \subset H^1(\Omega_L)$ is the finite element subspace with periodic boundary conditions in the expanding $\boldsymbol{x}$-directions and zero Dirichlet conditions in the $\boldsymbol{y}$-direction using the same polynomial order as $V_h(\Omega_L)$ on $\mathcal{T}_h$. Let $\tilde{\boldsymbol{\psi}}$ be the vector of coefficients of ψ_h and $\boldsymbol{\psi} \in \mathbb{R}^n$ its periodic extension to Ω_L. Then, using the shifted operator $\boldsymbol{A}_\sigma = \boldsymbol{A} - \sigma\boldsymbol{B}$ in the IPM results in convergence rates independently of L, as shown in Chapter 3.

Another class of eigensolvers interprets the EVP as an energy minimization problem that tries to minimize the Rayleigh quotient $R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}) := (\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})/(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x})$ since

$$\lambda_h^{(1)} = \min_{\substack{S\leq\mathbb{R}^n \\ |S|=1}} \max_{\substack{\boldsymbol{x}\in S \\ \boldsymbol{x}\neq\boldsymbol{0}}} R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}). \tag{4.13}$$

Taking the gradient, $\nabla R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}) = \frac{2}{\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x}}(\boldsymbol{A}\boldsymbol{x} - R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x})\boldsymbol{B}\boldsymbol{x})$, allows us to formulate the steepest descent method as

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \tau_k \nabla R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_k), \tag{4.14}$$

where the (adaptive) stepsize $\tau_k \in \mathbb{R}$ needs to be determined. Moving the scalar factor $2/(\boldsymbol{x}^T\boldsymbol{B}\boldsymbol{x})$ from $\nabla R_{\boldsymbol{A},\boldsymbol{B}}$ into τ_k, one can observe that the descent direction in the k-th step is proportional to the *spectral residual* $\boldsymbol{r}_k(\boldsymbol{x}) := \boldsymbol{A}\boldsymbol{x}_k - R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_k)\boldsymbol{B}\boldsymbol{x}_k$. To improve the convergence rate [31], we can also apply a preconditioner $\boldsymbol{P} \in \mathbb{R}^{n\times n}$ and use $\boldsymbol{w}_k = \boldsymbol{P}\boldsymbol{r}_k$ as the preconditioned search direction. Using the common choice of $\boldsymbol{P} = \boldsymbol{A}^{-1}$ and $\tau_k = 1$, the iteration $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \tau_k\boldsymbol{A}^{-1}\boldsymbol{r}_k$ reduces, up to normalization, back to the IPM. One might apply line search methods to find the optimal τ_k in each step. However, a common strategy is to use the Rayleigh–Ritz procedure [31, p39]

Algorithm 2 Inexact SI-Preconditioned LOPCG

Require: $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{n\times n}$; $\texttt{TOL}_\text{i}, \texttt{TOL}_\text{o} > 0$; $k_\text{max} \in \mathbb{N}$; vectors $\boldsymbol{x}_0, \boldsymbol{x}_1 \in \mathbb{R}^n$; shift $\sigma \in \mathbb{R}$

1: **function** SI-LOPCG
2: Initialize $k := 1$, $\boldsymbol{r}_{\text{o},k} := \boldsymbol{A}\boldsymbol{x}_k - R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_k)\boldsymbol{B}\boldsymbol{x}_k$, $\boldsymbol{A}_\sigma := \boldsymbol{A} - \sigma\boldsymbol{B}$
3: **while** $\|\boldsymbol{r}_{\text{o},k}\|_2 \leq \texttt{TOL}_\text{o}$ and $k \leq k_\text{max}$ **do** ▷ outer loop
4: Solve $\boldsymbol{A}_\sigma \boldsymbol{w}_k = \boldsymbol{r}_{\text{o},k}$ inexactly s.t. $\|\boldsymbol{A}_\sigma \boldsymbol{w}_k - \boldsymbol{r}_{\text{o},k}\|_2 \leq \texttt{TOL}_\text{i}$ ▷ inner loop
5: $S_k := \text{span}\{\boldsymbol{w}_k, \boldsymbol{x}_k, \boldsymbol{x}_{k-1}\}$
6: $\tilde{\boldsymbol{x}}_{k+1} := \arg\min_{\boldsymbol{y}\in S_k \setminus \{\boldsymbol{0}\}} R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{y})$ ▷ via orthogonalized Rayleigh–Ritz
7: Normalize $\boldsymbol{x}_{k+1} := \tilde{\boldsymbol{x}}_{k+1}/(\tilde{\boldsymbol{x}}_{k+1}^T \boldsymbol{B}\tilde{\boldsymbol{x}}_{k+1})^{1/2}$, set $k := k+1$, update $\boldsymbol{r}_{\text{o},k}$
8: **end while**
9: $\lambda_k := R_{\boldsymbol{A},\boldsymbol{B}}(\boldsymbol{x}_k)$
10: **return** eigenpair approximation $(\lambda_k, \boldsymbol{x}_k)$
11: **end function**

within the space $\tilde{S}_k := \text{span}\{\boldsymbol{w}_k, \boldsymbol{x}_k\}$ by solving a two-dimensional eigenvalue problem. The LOPCG [167] method adds another previous iterate $\boldsymbol{x}_{k-1}$ into the search space $S_k := \text{span}\{\boldsymbol{w}_k, \boldsymbol{x}_k, \boldsymbol{x}_{k-1}\}$ and finds the element with minimal $R_{\boldsymbol{A},\boldsymbol{B}}$ per iteration within the whole subspace S_k. As for the IPM, we also use the quasi-optimal shift-and-invert (SI) preconditioner $\boldsymbol{P} := \boldsymbol{A}_\sigma^{-1} = (\boldsymbol{A} - \sigma\boldsymbol{B})^{-1}$. This SI-preconditioner will also work for other gradient-based eigenvalue solvers.

However, applying the eigen-preconditioner $\boldsymbol{P}$ in practice amounts to solving a large and sparse linear system, which requires using iterative, inexact methods when n is large.

4.2.1 Inexact Inner-Outer Eigenvalue Algorithms

Since we deal with PDE-based problems on meshes, applying the eigenvalue preconditioner for $\boldsymbol{w}_k = \boldsymbol{A}_\sigma^{-1}\boldsymbol{r}$ uses an iterative method. This strategy leads to an *inner-outer* [114, 115] eigenvalue solver since inner iterations are needed to solve the linear systems for each outer eigenvalue iteration inexactly. As $\boldsymbol{A}_\sigma$ is symmetric positive-definite, the CG method is the preferred method. Defining an outer and an inner tolerance $0 < \texttt{TOL}_\text{o}, \texttt{TOL}_\text{i} \in \mathbb{R}$, the *inexact* SI-LOPCG (Algorithm 2) results. As the algorithm is a generalization of the plain inexact IPM, choosing $\texttt{TOL}_\text{i}$ sufficiently small ensures convergence [114, Cor. 3.2] with the usual assumptions.

The algorithm is similar to Newton–Krylov-type methods [166] for nonlinear equations, which fall back to the Rayleigh quotient iteration (RQI) for the EVP case. Although they have faster convergence rates, converging to the first eigenpair is not generally guaranteed [31, p53], so we stick with first-order methods.

4.2.2 Two-Level Domain Decomposition

In the inner linear system loop (line 4 of Algorithm 2), a system of the form $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ with a sparse matrix $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ and a given vector $\boldsymbol{b} \in \mathbb{R}^n$ ($\boldsymbol{A} \mapsto \boldsymbol{A}_\sigma, \boldsymbol{b} \mapsto \boldsymbol{r}_{\text{o},k}$

in our specific case) must be solved. Let $\mathcal{N}$ be the set of degrees of freedom, i.e., $|\mathcal{N}| = n$, and define an (overlapping) decomposition into $N \ll n$ subsets, $\{\mathcal{N}_i\}_{i=1}^N$, such that $\mathcal{N} = \cup_{i=1}^N \mathcal{N}_i$. This algebraic decomposition represents the geometric domain decomposition of Ω_L into subdomains. For all $\mathcal{N}_i$, there are associated *restriction* matrices $\boldsymbol{R}_i \in \{0,1\}^{|\mathcal{N}_i| \times |\mathcal{N}|}$, such that $\boldsymbol{R}_i \boldsymbol{x}$ restricts $\boldsymbol{x}$ to the subdomain $\mathcal{N}_i$. The transpose $\boldsymbol{R}_i^T \in \{0,1\}^{|\mathcal{N}| \times |\mathcal{N}_i|}$ is called the *extension* matrix. For the overlapping case with $\mathcal{N}_i \cap \mathcal{N}_j \neq \emptyset$ for some i, j, we can also define the diagonal *partition of unity* (PU) matrices $\boldsymbol{D}_i \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$, such that $\boldsymbol{x} = \sum_{i=1}^N \boldsymbol{R}_i^T \boldsymbol{D}_i \boldsymbol{R}_i \boldsymbol{x}$ for all $\boldsymbol{x}$ holds. An easy choice [98, p12] is to set $(\boldsymbol{D}_i)_{kk} := 1/|\mathcal{M}_k|$ where $\mathcal{M}_k := \{1 \leq i \leq N | k \in \mathcal{N}_i\}$ is the set of subdomains that contain the degree of freedom k.

Then, we can define the one-level *Additive Schwarz* [103] (AS) and the *restricted Additive Schwarz* [54] (RAS) preconditioners as

$$\boldsymbol{M}_{\mathrm{AS},1}^{-1} = \sum_{i=1}^N \boldsymbol{R}_i^T \boldsymbol{A}_i^{-1} \boldsymbol{R}_i, \quad \boldsymbol{M}_{\mathrm{RAS},1}^{-1} = \sum_{i=1}^N \boldsymbol{R}_i^T \boldsymbol{D}_i \boldsymbol{A}_i^{-1} \boldsymbol{R}_i, \tag{4.15}$$

where $\boldsymbol{A}_i := \boldsymbol{R}_i \boldsymbol{A} \boldsymbol{R}_i^T \in \mathbb{R}^{|\mathcal{N}_i| \times |\mathcal{N}_i|}$ are the subdomain coefficient matrices. We can formulate the *stationary RAS method* (i.e., a preconditioned fixed point iteration [98, p13]) to solve the linear system directly using

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \boldsymbol{M}_{\mathrm{RAS},1}^{-1} \boldsymbol{r}_k, \quad \boldsymbol{r}_k := \boldsymbol{b} - \boldsymbol{A}\boldsymbol{x}_k. \tag{4.16}$$

The RAS algorithm Eq. (4.16) moves away from the classical *"solve left, then use the data to solve right"*-paradigm (i.e., called multiplicative) but enables the parallel solution of each smaller subproblem, which is the deciding advantage of these additive methods concerning parallel computing.

Using Eq. (4.16) directly to solve QOSI-preconditioned inner systems would not result in a scalable method since the condition number of $\boldsymbol{M}_{\mathrm{RAS},1}^{-1} \boldsymbol{A}_\sigma$ will explode for $L \to \infty$. Thus, we will propose now a two-level modification to the AS/RAS-preconditioners that allows for full L-robustness. In general, the second level is usually incorporated via a coarse subspace $S_0 \leq \mathbb{R}^n$ with dimension $0 \leq n_0 \ll n$ that has an associated matrix $\boldsymbol{R}_0^T \in \mathbb{R}^{n \times n_0}$ (which we will specify in Section 4.2.3) such that the columns of $\boldsymbol{R}_0^T$ are linearly independent and span S_0. We can then define $\boldsymbol{A}_0 := \boldsymbol{R}_0 \boldsymbol{A} \boldsymbol{R}_0^T$ in the usual notation and include this *coarse correction* into both preconditioners from Eq. (4.15) in an additive fashion as

$$\boldsymbol{M}_{\star,2}^{-1} := \boldsymbol{R}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{R}_0 + \boldsymbol{M}_{\star,1}^{-1}, \tag{4.17}$$

where $\star$ stands for either AS or RAS. The coarse level can also be included in the stationary RAS iteration in a multiplicative fashion [124, p302] as

$$\begin{aligned} \boldsymbol{x}_{k+1/2} &= \boldsymbol{x}_k + \boldsymbol{M}_{\mathrm{RAS},1}^{-1} \boldsymbol{r}_k, \\ \boldsymbol{x}_{k+1} &= \boldsymbol{x}_{k+1/2} + \boldsymbol{R}_0^T \boldsymbol{A}_0^{-1} \boldsymbol{R}_0 \boldsymbol{r}_{k+1/2}. \end{aligned} \tag{4.18}$$

However, using the preconditioners directly in Krylov solvers is beneficial [89, p125], i.e., using the symmetric $\boldsymbol{M}_{\mathrm{AS},2}^{-1}$ for the CG and the unsymmetric $\boldsymbol{M}_{\mathrm{RAS},2}^{-1}$ for the GMRES method. With that description, the coarse space S_0 or its basis as columns of $\boldsymbol{R}_0^T$ remains to be specified.

4.2.3 PerFact: A Periodic Spectral Coarse Space Based on Asymptotic Factorization

For the classical Poisson problem, the usual motivation for coarse spaces is the lack of global information exchange [98, p103]. Decomposing a domain Ω_L into several subdomains necessarily leads to the emergence of tiny inner subdomains that are very far from the boundary in the subdomain connectivity graph. These inner subdomains then need a lot of DD iterations until the boundary data is propagated.

Although the initial problem (4.1) with positive V is spectrally equivalent to Poisson's problem, the situation is different in our setup. The focus is on the geometrical expansion of Ω_L with fixed subdomain and discretization size. So, choosing $N \sim L$ would not increase the distance of subdomains to the boundary and would suggest no need for a coarse space using the classical argument. However, the shifting strategy changes the spectrum of $\boldsymbol{B}^{-1}\boldsymbol{A}_\sigma$ by shifting it closer to the origin. Thus, the asymptotic loss of coercivity is the mechanism behind the need for a coarse space in our setup.

However, the usual rule to include *slow modes* in V_0 still applies, although low-energy modes are the more precise terminology. Moreover, surprisingly, these modes are already at hand. Reconsider the factorization of the ground-state solution ϕ_L of Eq. (4.1) into the unit cell solution ψ of Eq. (4.4) plus a remainder u as

$$\phi_L = \psi \cdot u, \quad \lambda_L = \sigma + \lambda_u \text{ where } \lambda_u \in \mathcal{O}(1/L^2), \tag{4.19}$$

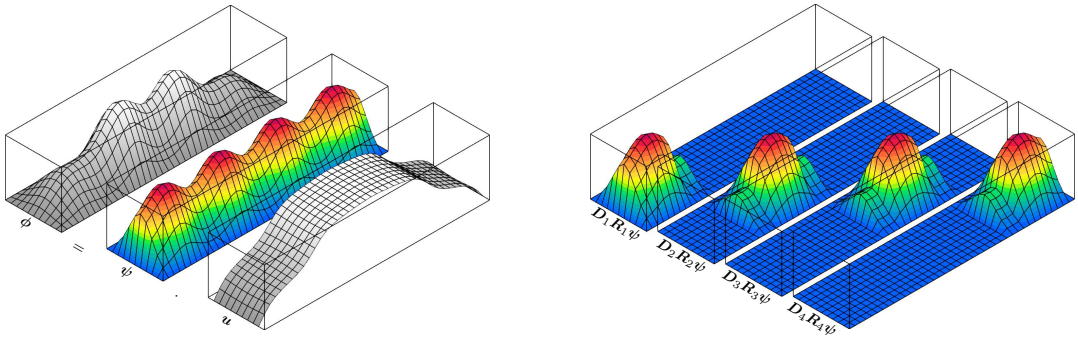
that was shown in Eq. (4.19) and Theorem 3.2. In Eq. (4.19), σ is the asymptotic limit of the desired first eigenvalue. Equivalently spoken, the function ψ corresponding to σ approaches more and more an eigenfunction of the shifted $\boldsymbol{A}_\sigma$ operator with zero eigenvalue – a zero energy mode. This observation motivates us to include exactly this function in the coarse space. We define the following with the coefficient vector $\boldsymbol{\psi} \in \mathbb{R}^n$ from the periodic unit cell problem (4.12).

Definition 4.1. The matrix

$$\boldsymbol{R}_0^T := \begin{bmatrix} \boldsymbol{D}_1\boldsymbol{R}_1\boldsymbol{\psi} & 0 & \cdots & 0 \\ 0 & \boldsymbol{D}_2\boldsymbol{R}_2\boldsymbol{\psi} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \boldsymbol{D}_N\boldsymbol{R}_N\boldsymbol{\psi} \end{bmatrix} \in \mathbb{R}^{n \times N}, \tag{4.20}$$

defines the components of the periodic factorization (PerFact) coarse space.

Remark 4.1. The structure of Eq. (4.20) resembles the shape of the classical Nicolaides coarse space where the local Laplacian kernels, i.e., the constant vectors $\mathbf{1}$, are replaced by their corresponding local kernels $\boldsymbol{\psi}$ of the shifted Schrödinger operator. While the constant vectors $\mathbf{1}$ might be able to capture global information exchange, we will show that the high-frequency features of $\boldsymbol{\psi}$ must also be present for the coarse space to work. Also note that Eq. (4.20) technically contains more basis functions



(a) Visualization of the factorization principle.

(b) Components of the PerFact coarse space.

Figure 4.2: Finite element representation of the **(a)** factorization principle from Eq. (4.19) and the **(b)** coarse space basis functions from Eq. (4.20) for the equal-weights partition of unity and an overlap region of two elements between subdomains.

than needed, as we will show in Section 4.3, since it also includes boundary-touching domains. However, we keep this structure for simplicity and consistency with the literature [98, p108].

In Fig. 4.2, we display the factorization principle Eq. (4.19) and the coarse space components Eq. (4.20) for an exemplary decomposition into four subdomains. Therein, the main advantages become visible – the vector $\boldsymbol{\psi}$ is already available from the computation of the QOSI shift (i.e., problem Eq. (4.4)) and the small dimension $n_0 = N$ since every subdomain only needs one basis function in the coarse space in contrast to other adaptive coarse space strategies. The PerFact coarse space is mainly used in the two-level additive Schwarz preconditioner within the CG method, but an application of the stationary two-level RAS algorithm also works, as we will see in Section 4.5.

4.3 Analysis of the Two-Level Additive Schwarz Preconditioner

In general, using a s.p.d. preconditioner $\boldsymbol{M}$ within the CG method to solve the shifted linear systems $\boldsymbol{A}_\sigma \boldsymbol{x} = \boldsymbol{b}$ leads to the convergence estimate (see, e.g., [253, Lem. C.10]) of

$$\|\boldsymbol{x} - \boldsymbol{x}_k\|_{\boldsymbol{A}_\sigma} \leq 2\left(\frac{\sqrt{\kappa(\boldsymbol{M}^{-1}\boldsymbol{A}_\sigma)} - 1}{\sqrt{\kappa(\boldsymbol{M}^{-1}\boldsymbol{A}_\sigma)} + 1}\right)^k \|\boldsymbol{x} - \boldsymbol{x}_0\|_{\boldsymbol{A}_\sigma}. \tag{4.21}$$

For increasing domain sizes L, applying the quasi-optimal shift-and-invert technique from Chapter 3 results in $\lambda_{\min}(\boldsymbol{A}_\sigma) \to 0$ as $L \to \infty$. However, since the domain expands anisotropically with constant ℓ, we have $\lambda_{\max}(\boldsymbol{A}_\sigma) \geq C > 0$ for all $L \to \infty$

by using the Laplacian EVP as a lower bound and following the argumentation of Section 4.1.2.

The behavior of the spectrum thus results in $\kappa(\boldsymbol{A}_\sigma) \to \infty$ for $L \to \infty$. Therefore, without a preconditioner, i.e., $\boldsymbol{M} = \boldsymbol{I}$, the number of iterations to decrease the relative error below a given tolerance will drastically increase for $L \to \infty$ in the linear solution step of Algorithm 2 (line 4). However, the two-level AS preconditioner $\boldsymbol{M}_{\mathrm{AS},2}^{-1}$ from Eq. (4.17) is L-robust, i.e., $\kappa(\boldsymbol{M}_{\mathrm{AS},2}^{-1}\boldsymbol{A}_\sigma) \leq C$ for all L, since we can control the lower part of the spectrum with the PerFact coarse space, as shown in the following.

4.3.1 Abstract Theory

After the conforming finite element discretization with $\overline{\Omega}_L = \bigcup_{\tau \in \mathcal{T}_h} \tau$ from Section 4.2, we operate in the finite-dimensional setting with $V_h(\Omega_L) \subset H_0^1(\Omega_L)$, where we now use the abbreviation $V_h := V_h(\Omega_L)$ for simplicity. In each step of the eigenvalue algorithm, we have to apply the preconditioner, which amounts to solving a source problem: Given $f_h \in V_h'$, find $u_h \in V_h$ such that

$$a_\sigma(u_h, v_h) = \langle f_h, v_h \rangle, \quad \forall v_h \in V_h. \tag{4.22}$$

The bilinear form $a_\sigma : V_h \times V_h \to \mathbb{R}$ in Eq. (4.22) corresponds to the shift-and-invert eigenvalue preconditioner matrix $\boldsymbol{A}_\sigma$ and is given by

$$a_\sigma(u, v) := (\nabla u, \nabla v)_{L^2(\Omega_L)} + (Vu, v)_{L^2(\Omega_L)} - \sigma(u, v)_{L^2(\Omega_L)}. \tag{4.23}$$

In order to apply a two-level additive Schwarz method, we decompose the domain Ω_L into a non-overlapping set of subdomains $\{\Omega_i'\}_{i=1}^N$ either by using the natural decomposition into $N = L^p$ repeating unit cells or with the help of an automatic graph partitioning software, e.g., METIS [161] or Scotch [75]. Then, let $\{\Omega_i\}_{i=1}^L$ with $\Omega_i' \subset \Omega_i$ be an overlapping domain decomposition by adding $\delta \geq 1$ layers of elements $\tau \in \mathcal{T}_h$ based on the element connectivity graph. Following the abstract theory, c.f., e.g., [98], we can define for $i = 1, \ldots, N$ the space of restrictions of functions $v \in V_h$ to a subdomain Ω_i as $V_h(\Omega_i) := \{v|_{\Omega_i} \mid v \in V_h\}$ and the space of $\overline{\Omega}_i$-supported functions as $V_{h,0}(\Omega_i) := \{v|_{\Omega_i} \mid v \in V_h,\ \mathrm{supp}(v) \subset \overline{\Omega}_i\}$. Then we have the restriction operators $r_i : V_h \to V_h(\Omega_i), r_i v = v|_{\Omega_i}$ and their adjoints, the extension-by-zero operators, $r_i^T : V_{h,0}(\Omega_i) \to V_h$, which extends functions by zero outside of Ω_i. For simplicity, we will sometimes interpret $V_{h,0}(\Omega_i)$ as a subspace of V_h leaving out the extension-by-zero operators.

For the two-level approach, we also define a coarse space $V_0 \subset V_h$ with corresponding natural embedding (i.e., the inclusion map that is the linear interpolation [180] in the present setup) $r_0^T : V_0 \to V_h$ and its adjoint $r_0 : V_h \to V_0$. The matrix form of the two-level additive Schwarz preconditioner for $\boldsymbol{A}_\sigma$ then reads

$$\boldsymbol{M}_{\mathrm{AS},2}^{-1} = \boldsymbol{R}_0^T \boldsymbol{A}_{\sigma,0}^{-1} \boldsymbol{R}_0 + \sum_{i=1}^N \boldsymbol{R}_i^T \boldsymbol{A}_{\sigma,i}^{-1} \boldsymbol{R}_i \text{ with } \boldsymbol{A}_{\sigma,i} = \boldsymbol{R}_i \boldsymbol{A}_\sigma \boldsymbol{R}_i^T \text{ and } \boldsymbol{A}_{\sigma,0} = \boldsymbol{R}_0 \boldsymbol{A}_\sigma \boldsymbol{R}_0^T, \tag{4.24}$$

with $\{\boldsymbol{R}_i\}_{i=1}^N$ and $\boldsymbol{R}_0$ denoting the matrix representations of $\{r_i\}_{i=1}^N$ and r_0 for the given finite element basis. The local matrices $\boldsymbol{A}_{\sigma,i}$ are invertible since the corresponding restricted bilinear forms $a_{\sigma,\Omega_i} : V_{h,0}(\Omega_i) \times V_{h,0}(\Omega_i)$ are positive definite as $a_{\sigma,\Omega_i}(u,v) := a_\sigma(r_i^T u, r_i^T v) > 0$ for all $u, v \in V_{h,0}(\Omega_i)$ and finite L.

Following [98, 235], we now recall the main ingredients of the abstract additive Schwarz theory. First, we have the following geometric definition.

Definition 4.2 (finite coloring [36, Def. 3.1])**.** The partition $\{\Omega_i'\}_{i=1}^N$ admits a finite coloring with $N_\mathrm{c} \in \mathbb{N}$ colors, $N_\mathrm{c} \leq N$, if there exists a map $c : \{1,\dots,N\} \to \{1,\dots,N_\mathrm{c}\}$ such that

$$i \neq j \wedge c(i) = c(j) \Rightarrow a_\sigma(r_i^T v_i, r_j^T v_j) = 0, \quad \forall v_i \in V_{h,0}(\Omega_i), v_j \in V_{h,0}(\Omega_j). \tag{4.25}$$

In practice, a low N_c is naturally given when aligning the decomposition with the period of V. The following notion is the critical ingredient to show the effectiveness of a coarse space by lower bounding the smallest eigenvalues of the preconditioned system.

Definition 4.3 (stable decomposition [235, Def. 2.7] [36, Def. 3.3])**.** Given a coarse space $V_0 \subset V_h$, local subspaces $\{V_{h,0}(\Omega_i)\}_{1\leq i\leq N}$ of V_h, and a constant $C_0 > 0$, a stable decomposition of $v \in V_h$ is a family of functions $\{v_i\}_{0\leq i\leq N}$ that satisfy $v = v_0 + \sum_{i=1}^N v_i$, $v_i \in V_{h,0}(\Omega_i)$, $v_0 \in V_0$, such that

$$a_\sigma(v_0, v_0) + \sum_{i=1}^N a_\sigma(v_i, v_i) \leq C_0\, a_\sigma(v, v). \tag{4.26}$$

Thus, according to [98, Cor. 5.12], we aim to find a stable decomposition for all $v \in V_h$ since it directly yields a condition number bound of $\boldsymbol{M}_{\mathrm{AS},2}^{-1}\boldsymbol{A}_\sigma$. To apply this abstract theory, we need further notation.

Definition 4.4 (partition of unity [98, Lem. 5.7])**.** For the overlapping decomposition $\{\Omega_i\}_{i=1}^N$ of Ω_L, there exists a family of partition of unity functions $\{\chi_i\}_{i=1}^N$ in $W^{1,\infty}(\Omega_L)$, such that $0 \leq \chi_i(\boldsymbol{z}) \leq 1$ for $\boldsymbol{z} \in \overline{\Omega}_L$, $\mathrm{supp}(\chi_i) \subset \overline{\Omega}_i$, and $\sum_{i=1}^N \chi_i(\boldsymbol{z}) = 1$ for all $\boldsymbol{z} \in \overline{\Omega}_L$.

Proof. See, e.g., [253, Lem. 3.4] with $\{\Omega_i\}_{i=1}^N$ satisfying the overlap and finite coloring assumptions. □

Remark 4.2. For the construction of the partition of unity, we use the finite element basis representation $v = \sum_{k=1}^n v_k N_k$ for any $v \in V_h$. With $\mathrm{dof}(\Omega_i)$ denoting the set of all internal degrees of freedoms for subdomain Ω_i, the most straightforward choice is the equal-weights partition of unity [235] where $\chi_i(v) := \sum_{k\in\mathrm{dof}(\Omega_i)} \frac{1}{\mu_k} v_k N_k$ with $\mu_k := |\{j \mid 1 \leq j \leq N, k \in \mathrm{dof}(\Omega_j)\}|$ denoting the number of subdomains for which k is an internal degree of freedom. Other choices are based on distance functions, e.g. [253, Lem. 3.4], and have the additional property that $\|\nabla\chi_i\|_\infty \leq C/\delta_i$, where δ_i is the minimal overlap thickness of the i-th subdomain. This distance-based partition of unity is favorable when the overlap varies.

In calculating the quasi-optimal shift σ, we already solved the eigenvalue problem on the unit cell Ω_1 with $\boldsymbol{x}$-periodic and $\boldsymbol{y}$-zero boundary conditions in Eq. (4.12), where we abbreviate its solution ψ_h form now on as ψ for simplicity. We also define $E_{\boldsymbol{x}}^{\#} : V_{h,\#}(\Omega_1) \to V_h$ as the periodic extension operator in the $\boldsymbol{x}$-direction. $E_{\boldsymbol{x}}^{\#}$ acts only on the degrees of freedom, such that the result is only periodic up to the boundary nodes of Ω_L, which are set to zero for the result to be in V_h. Although we do not solve local eigenvalue problems in practice since ψ is already computed, we can still analyze the method in the context of spectral coarse spaces when we take care of the periodicity by modifying the decomposition.

4.3.2 Aligning the Decomposition

Recall from Eq. (4.2) that the domain Ω_L is an abstract box. Naturally, Ω_L is split into L^p cells, so every cell can be identified with an integer vector $\boldsymbol{i} \in \{1, \dots, L\}^p$ when considering the distance to the unit cell. We map this vector translation to a one-dimensional index via the bijective map $n : \{1, \dots, L\}^p \to \{1, \dots, L^p\}$ where $n(\boldsymbol{i}) := 1 + \sum_{j=1}^{p} L^{j-1}(\boldsymbol{i}_j - 1)$. The corresponding inverse map $\boldsymbol{n}^{-1} : \{1, \dots, L^p\} \to \{1, \dots, L\}^p$ can then be used to define the *periodic decomposition* $\{\Omega_i^{\#}\}_{i=1}^{L^p}$, where each periodic cell is given by

$$\Omega_i^{\#} := \bigtimes_{j=1}^{p} \left([\boldsymbol{n}^{-1}(i)]_j - 1, [\boldsymbol{n}^{-1}(i)]_j \right) \times (0, \ell)^q. \tag{4.27}$$

Every subdomain Ω_i belongs to a periodic neighborhood, defined by the following.

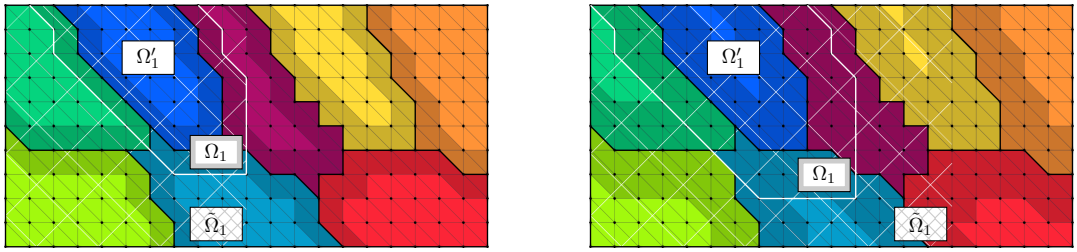
Definition 4.5 (periodic neighborhood)**.** Let $\Omega_i \in \{\Omega_i\}_{i=1}^{N}$ be any subdomain. Then, the corresponding periodic neighborhood $\tilde{\Omega}_i$ of Ω_i is defined by

$$\tilde{\Omega}_i := \bigcup_{i \in \mathcal{I}_i^{\#,\mathrm{n}}} \Omega_i^{\#} \text{ where } \mathcal{I}_i^{\#,\mathrm{n}} := \left\{ j \in \{1, \dots, L^p\} \;\middle|\; \Omega_i \cap \Omega_j^{\#} \neq \emptyset \right\}. \tag{4.28}$$

The Fig. 4.3 visualizes the relation between Ω_i, Ω_i', and $\tilde{\Omega}_i$. A collection of periodic neighborhoods $\{\tilde{\Omega}_i\}_{i=1}^{N}$ is itself again an overlapping decomposition of Ω_L, where the decomposition reflects the periodicity of the potential V. This fact will be helpful later on and allows us to define the subspaces $V_{h,\#}(\tilde{\Omega}_i) \subset V_h(\tilde{\Omega}_i)$ with functions that are periodic on the $\boldsymbol{x}$-boundary of $\tilde{\Omega}_i$. The periodic neighborhood decomposition also induces their corresponding restrictions $\tilde{r}_i : V_h \to V_h(\tilde{\Omega}_i)$ and partitions of unity $\tilde{\chi}_i$, defined by the zero-extension of χ_i from Ω_i to $\tilde{\Omega}_i$. The set $\{\tilde{\Omega}_i\}_{i=1}^{N}$ gives rise to a multiplicity constant in the following.

Definition 4.6 (periodic neighborhood intersection multiplicity)**.** Let $\tilde{k}_0$ be the maximum number of periodic neighborhoods to which one periodic cell from $\{\Omega_i^{\#}\}_{i=1}^{L^p}$ can belong, i.e., $\tilde{k}_0 := \max_{i \in \{1,\dots,N\}} |\{j \in \{1, \dots, L^p\} \mid \tilde{\Omega}_i \cap \Omega_j^{\#} \neq \emptyset\}|$.

Similar to the classical Nicolaides coarse space for the Laplace problem [99], we only include subdomains away from the boundary in the coarse space due to technical



(a) $\delta = 1$. (b) $\delta = 2$.

Figure 4.3: Sketch of the non-overlapping $\{\Omega_i'\}_{i=1}^8$ (black border), overlapping $\{\Omega_i\}_{i=1}^8$ (white border), and periodic neighborhood decomposition $\{\tilde{\Omega}_i\}_{i=1}^8$ (cross-hatch) of $\Omega_4 := (0,4) \times (0,2)$ for an overlap (dark shades) of **(a)** $\delta = 1$ and **(b)** $\delta = 2$ layers of elements. An increase of the periodic neighborhood from $\tilde{\Omega}_1 = (0,2) \times (0,2)$ for $\delta = 1$ to $\tilde{\Omega}_1 = (0,3) \times (0,2)$ for $\delta = 2$ can be observed.

reasons, such as non-matching boundary conditions. In the present setup, only the boundaries in the expanding $\boldsymbol{x}$-direction need particular focus. Thus, let us define the set of all $\boldsymbol{x}$-boundary domains as $\mathcal{I}^{\mathrm{b}} := \{i \in 1, \dots, N \mid \partial\Omega_i \cap (\{0, L\}^p \times (0, \ell)^q) \neq \emptyset\}$. This induces the set of all interior domains as $\mathcal{I}^{\mathrm{i}} := \{1, \dots, N\} \setminus \mathcal{I}^{\mathrm{b}}$. Let us also define a set of domains $\tilde{\mathcal{I}}^{\mathrm{b}} := \{i \in 1, \dots, N \mid \partial\tilde{\Omega}_i \cap (\{0, L\}^p \times (0, \ell)^q) \neq \emptyset\}$, whose periodic neighborhoods touch the $\boldsymbol{x}$-boundary. The other periodic neighborhoods are grouped into $\tilde{\mathcal{I}}^{\mathrm{i}} := \{1, \dots, N\} \setminus \tilde{\mathcal{I}}^{\mathrm{b}}$. Based on this period-aligned splitting, we define the periodic coarse space.

Definition 4.7 (PerFact coarse space)**.** With ψ from Eq. (4.12), the space $V_0 := \operatorname{span}\left(\{\chi_i E_{\boldsymbol{x}}^{\#} \psi\}_{i \in \tilde{\mathcal{I}}^{\mathrm{i}}}\right)$ is called the periodic factorization (PerFact) coarse space.

Remark 4.3. Note that the Definition 4.7 and the Definition 4.1 only differ for $\boldsymbol{x}$-boundary touching subdomains, which are L-asymptotically irrelevant.

For the analysis, we also need the notion of cell symmetry.

Definition 4.8 ($\boldsymbol{x}$-cell-symmetric mesh)**.** Let $\mathcal{T}_h$ be a triangulation, $\bar{x}_i$ the i-th component of its center of mass $\bar{\boldsymbol{x}}$, and $R_{\bar{x}_i}$ the reflection across the plane $P_{\bar{x}_i} = \left\{\boldsymbol{z} \in \mathbb{R}^d \;\middle|\; \langle e_i, \boldsymbol{z}\rangle = \bar{x}_i\right\}$. We call $\mathcal{T}_h$ to be $\boldsymbol{x}$-symmetric if for all expanding directions with $i \in \{1, \dots, p\}$, $\tau \in \mathcal{T}_h \Rightarrow R_{\bar{x}_i}\tau \in \mathcal{T}_h$.

Definition 4.9 ($\boldsymbol{x}$-cell-symmetric potential)**.** A potential V is said to be $\boldsymbol{x}$-cell-symmetric for an $\tilde{\Omega}_i$ if for all x_i-dimensions with $i \in \{1, \dots, p\}$

$$V(x_1, \dots, x_i, \dots, x_p, \boldsymbol{y}) = V(x_1, \dots, 2\bar{x}_i - x_i, \dots, x_p, \boldsymbol{y}) \quad \text{for a.e. } (\boldsymbol{x}, \boldsymbol{y}) \in \tilde{\Omega}_i. \tag{4.29}$$

4.3.3 A Condition Number Bound for Cell-Symmetric Potentials

We are now prepared to analyze the condition number of the preconditioned system. As we operate near the *edge of coercivity*, meaning that we shifted the operator by

σ, which is only a little less than the smallest eigenvalue, we must be cautious with the analysis. In particular, we must ensure that local Neumann problems remain positive semidefinite for the theory of spectral coarse spaces to hold. Thus, we base the analysis on the following assumptions.

(B3) The triangulation of all periodic neighborhoods $\tilde{\Omega}_i$ is $\boldsymbol{x}$-cell-symmetric.

(B4) The potential V is $\boldsymbol{x}$-cell-symmetric for all $\tilde{\Omega}_i$.

Although these assumptions seem restricted, the method still works in the general case, as shown numerically in Section 4.5. To apply the theory of spectral coarse spaces, we show that $E_{\boldsymbol{x}}^{\#}\psi$, defined by Eq. (4.12), is a solution to a generalized eigenvalue problem (GEVP) of a particular structure. In the usual setting, the GEVPs are formulated on overlapping subdomains Ω_i or inside the overlapping zone. In our setting, however, the components $E_{\boldsymbol{x}}^{\#}\psi$ from Definition 4.7 are periodic. This property can be used if we formulate the GEVPs on the periodic neighborhood $\tilde{\Omega}_i$.

Definition 4.10 (GEVPs)**.** For all interior periodic neighborhoods $\tilde{\Omega}_i$ with $i \in \tilde{\mathcal{I}}^{\mathrm{i}}$, we define the following generalized eigenvalue problems on W_i: Find $(\tilde{p}_i^{(k)}, \tilde{\lambda}_i^{(k)}) \in (W_i \setminus \{0\}) \times \mathbb{R}$, such that

$$\tilde{a}_{\sigma,\tilde{\Omega}_i}(\tilde{p}_i^{(k)}, v) = \tilde{\lambda}_i^{(k)} \tilde{b}_{\tilde{\Omega}_i}(\tilde{p}_i^{(k)}, v), \quad \forall v \in W_i, \tag{4.30}$$

where we set $\tilde{b}_{\tilde{\Omega}_i}(u, v) := \tilde{a}_{0,\tilde{\Omega}_i}(\tilde{\chi}_i u, \tilde{\chi}_i v)$. The default case of $W_i = V_h(\tilde{\Omega}_i)$ is called the Neumann GEVP (NGEVP), and the case of $W_i = V_{h,\#}(\tilde{\Omega}_i)$ is called the periodic GEVP (PGEVP), whose eigenpair is then denoted by $(\tilde{p}_{i,\#}^{(k)}, \tilde{\lambda}_{i,\#}^{(k)})$.

Remark 4.4. In general, there is some flexibility in choosing the bilinear form $\tilde{b}_{\tilde{\Omega}_i}$ in Eq. (4.30), as discussed in [36, 45]. Although the resulting eigenfunctions would be changed for different $\tilde{b}_{\tilde{\Omega}_i}$, we are only interested in the kernel of $\tilde{a}_{\sigma,\tilde{\Omega}_i}$, which is unaffected by the specific choice of $\tilde{b}_{\tilde{\Omega}_i}$ as long as the GEVPs are non-defective. This provides some freedom in carrying out the analysis.

Remark 4.5. In contrast to the classical approaches in coarse spaces, we do not use the same generating bilinear form to define the left- and the right-hand side of the GEVP. In particular, only the left side of Eq. (4.30) has a negative shift term in $\tilde{a}_{\sigma,\tilde{\Omega}_i}$, while the right-hand side is generated without shift by $\tilde{a}_{0,\tilde{\Omega}_i}$. This ensures that $\tilde{b}_\tau(u|_\tau, u|_\tau) \geq 0 \ \forall u \in V_h(\tilde{\Omega}_i)$, which guarantees the positive semidefiniteness of $\tilde{b}_{\tilde{\Omega}_i}$, which is a crucial requirement for the theory. A similar idea finds application in the context of non-self-adjoint or indefinite problems, e.g., in the $\mathcal{H}$-GenEO approach of [45].

We will now show one of the crucial observations: the PGEVP and the NGEVP have the same first eigenpair under the assumptions (B3) and (B4). To formalize that result, we need the following notion of reflection.

Definition 4.11 (reflection operator)**.** Let $\tilde{\Omega}_i$ be an $\boldsymbol{x}$-symmetric triangulation. For any function $v \in V_h(\tilde{\Omega}_i)$ with $m = |V_h(\tilde{\Omega}_i)|$ DOFs, we define the reflection $R_{\bar{x}_i} : V_h(\tilde{\Omega}_i) \to V_h(\tilde{\Omega}_i)$ across the $P_{\bar{x}_i}$-plane using the finite element basis representation $v = \sum_{k=1}^m v_k N_k$ as

$$v \mapsto R_{\bar{x}_i} v := \sum_{k=1}^{m} \left(\sum_{l=1}^{m} R_{kl} v_l \right) N_k, \tag{4.31}$$

where the permutation matrix $\boldsymbol{R} := (R_{kl})_{1 \le k,l \le m}$ (after reordering) has the form

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{I}_{m_1} \otimes \boldsymbol{S} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{I}_{m_2} \end{bmatrix}, \text{ with the pairwise swap matrix } \boldsymbol{S} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{4.32}$$

where all m nodes are split into m_2 on the reflection plane and $2m_1$ remaining nodes.

Lemma 4.1. *Assume (B3) and (B4), then the unique lowest eigenpair of the NGEVPs and PGEVPs from Definition 4.10 is given by* $(\tilde{p}_i^{(1)}, \tilde{\lambda}_i^{(1)}) = (\tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi, 0)$.

Proof. Let $i \in \tilde{\mathcal{I}}^{\mathrm{i}}$. Assume that $p^{(1)}$ is the first eigenfunction of the homogeneous EVP with the standard $L^2(V_h(\tilde{\Omega}_i))$-inner product, meaning

$$p^{(1)} \in \operatorname*{arg\,min}_{u \in V_h(\tilde{\Omega}_i)} \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(u,u)}{(u,u)_{L^2(V_h(\tilde{\Omega}_i))}}. \tag{4.33}$$

By [127, Thm. 8.38], $p^{(1)}$ is simple and strictly positive (by fixing the sign). Now fix an arbitrary x_j-dimension with $j \in \{1, \dots, p\}$ and consider the reflection $q^{(1)} := R_{\bar{x}_j} p^{(1)}$ of $p^{(1)}$ across the plane $P_{\bar{x}_j}$. Regarding the basis coefficient vector $\boldsymbol{p}$ of $p^{(1)}$, this means $\boldsymbol{q} = \boldsymbol{R}\boldsymbol{p}$ in the sense of Definition 4.11. Now, two things can happen.
Case 1: $\boldsymbol{q} \in \operatorname{span} \boldsymbol{p}$ In this case, there exists an $\alpha \in \mathbb{R}$ such that $\boldsymbol{p}$ is an eigenvector with $\boldsymbol{q} := \boldsymbol{R}\boldsymbol{p} = \alpha \boldsymbol{p}$. Since the reflection matrix $\boldsymbol{R}$ is a permutation matrix and thus an orthogonal matrix, it only has eigenvalues $\alpha \in \{-1, 1\}$, such that there are again two cases. For $\alpha = 1$, we have that $p^{(1)}$ is symmetric w.r.t. the plane $P_{\bar{x}_j}$ since it is invariant under the reflection. The other case of $\alpha = -1$ is impossible since all coefficients in $\boldsymbol{p}$ are strictly positive, and a permutation matrix $\boldsymbol{R}$ can not change this property for the equality $\boldsymbol{p} = -\boldsymbol{R}\boldsymbol{p}$ to hold.
Case 2: $\boldsymbol{q} \notin \operatorname{span} \boldsymbol{p}$ In that case, we observe that $q^{(1)}$ and $p^{(1)}$ have the same Rayleigh quotient since

$$\frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(q^{(1)}, q^{(1)})}{(q^{(1)}, q^{(1)})_{L^2(V_h(\tilde{\Omega}_i))}} = \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(R_{\bar{x}_j} p^{(1)}, R_{\bar{x}_j} p^{(1)})}{(R_{\bar{x}_j} p^{(1)}, R_{\bar{x}_j} p^{(1)})_{L^2(V_h(\tilde{\Omega}_i))}} = \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(p^{(1)}, p^{(1)})}{(p^{(1)}, p^{(1)})_{L^2(V_h(\tilde{\Omega}_i))}}, \tag{4.34}$$

in which we used $\|R_{\bar{x}_j} p^{(1)}\|^2_{L^2(\tilde{\Omega}_i)} = \|p^{(1)}\|^2_{L^2(\tilde{\Omega}_i)}$ by the orthogonality of the reflection $R_{\bar{x}_j}$ and $\tilde{a}_{\sigma,\tilde{\Omega}_i}(R_{\bar{x}_j} p^{(1)}, R_{\bar{x}_j} p^{(1)}) = \tilde{a}_{\sigma,\tilde{\Omega}_i}(p^{(1)}, p^{(1)})$ by the symmetry of the mesh $\tilde{\Omega}_i$ and the potential V w.r.t. to the $P_{\bar{x}_j}$-plane, (B3) and (B4). However, Eq. (4.34) would imply that the first eigenspace is two-dimensional since $\boldsymbol{q}$ is not in the span $\boldsymbol{p}$, although it has a minimal Rayleigh quotient, which is a contradiction.

Thus, all in all, $p^{(1)}$ must be symmetric w.r.t. the $P_{\bar{x}_j}$-plane. Since the arguments apply for all expanding directions, the first eigenfunction $p^{(1)}$ is $\boldsymbol{x}$-symmetric, i.e., $p^{(1)} = R_{\bar{x}_j} p^{(1)}$ for all $j \in \{1, \dots, p\}$, implying that $\tilde{b}_{\tilde{\Omega}_i}(R_{\bar{x}_j} p^{(1)}, R_{\bar{x}_j} p^{(1)}) = \tilde{b}_{\tilde{\Omega}_i}(p^{(1)}, p^{(1)})$. Symmetry also yields periodicity on the $\boldsymbol{x}$-boundaries resulting in $p^{(1)} \in V_{h,\#}(\tilde{\Omega}_i)$. Using the min-max principle for the GEVPs, we thus have

$$\min_{u \in V_h(\tilde{\Omega}_i)} \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(u,u)}{\tilde{b}_{\tilde{\Omega}_i}(u,u)} = \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(p^{(1)},p^{(1)})}{\tilde{b}_{\tilde{\Omega}_i}(p^{(1)},p^{(1)})} \geq \min_{u \in V_{h,\#}(\tilde{\Omega}_i)} \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(u,u)}{\tilde{b}_{\tilde{\Omega}_i}(u,u)} = 0, \tag{4.35}$$

where the last equality holds by definition of the shift σ. However, due to the subset property, $V_{h,\#}(\tilde{\Omega}_i) \subset V_h(\tilde{\Omega}_i)$, we also have

$$\min_{u \in V_h(\tilde{\Omega}_i)} \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(u,u)}{\tilde{b}_{\tilde{\Omega}_i}(u,u)} \leq \min_{u \in V_{h,\#}(\tilde{\Omega}_i)} \frac{\tilde{a}_{\sigma,\tilde{\Omega}_i}(u,u)}{\tilde{b}_{\tilde{\Omega}_i}(u,u)} = 0. \tag{4.36}$$

Thus, the NGEVPs and PGEVPs have the same first eigenpair $(\tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi, 0)$. □

Corollary 4.1. *Assume (B3) and (B4), then the bilinear form* $\tilde{a}_{\sigma,\tilde{\Omega}_i} : V_h(\tilde{\Omega}_i) \times V_h(\tilde{\Omega}_i)$ *is positive semidefinite since its lowest eigenvalue is* 0.

We verify the usual ingredients to show the C_0-stability in the following. Since $\tilde{a}_{\sigma,\tilde{\Omega}_i}$ is positive semidefinite on $V_h(\tilde{\Omega}_i)$ by Corollary 4.1 while $\tilde{b}_{\tilde{\Omega}_i}$ is positive semidefinite as the sum of elementwise non-negative integrals (see Definition 4.10), we define the two induced seminorms $|v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}} := (\tilde{a}_{\sigma,\tilde{\Omega}_i}(v,v))^{1/2}$ and $|v|_{\tilde{b}_{\tilde{\Omega}_i}} := (\tilde{b}_{\tilde{\Omega}_i}(v,v))^{1/2}$. We then have the following result.

Lemma 4.2 (SPSD splitting [36, Def. 3.7])**.** *Assume (B3) and (B4), then*

$$\sum_{i=1}^{N} |\tilde{r}_i v|^2_{\tilde{a}_{\sigma,\tilde{\Omega}_i}} \leq \tilde{k}_0 \|v\|^2_{a_\sigma}, \quad \forall v \in V_h. \tag{4.37}$$

Proof. Let $v \in V_h$. We use $\tilde{\Omega}_i := \bigcup_{i \in \mathcal{I}_i^{\#,\mathrm{n}}} \Omega_i^{\#}$ from Definition 4.5 to obtain

$$\sum_{i=1}^{N} |\tilde{r}_i v|^2_{\tilde{a}_{\sigma,\tilde{\Omega}_i}} = \sum_{i=1}^{N} \sum_{\{j \mid \Omega_j^{\#} \subset \tilde{\Omega}_i\}} \tilde{a}_{\sigma,\Omega_j^{\#}}(v|_{\Omega_j^{\#}}, v|_{\Omega_j^{\#}}) = \sum_{j=1}^{L^p} \sum_{\{i \mid \Omega_j^{\#} \subset \tilde{\Omega}_i\}} \tilde{a}_{\sigma,\Omega_j^{\#}}(v|_{\Omega_j^{\#}}, v|_{\Omega_j^{\#}}). \tag{4.38}$$

Using that each $\Omega_j^{\#}$ is contained in at most $\tilde{k}_0$ periodic neighborhoods $\tilde{\Omega}_i$ and the positive semidefiniteness of $\tilde{a}_{\sigma,\Omega_j^{\#}}$ on $V_h(\Omega_j^{\#})$ (see the proof of Lemma 4.1) yields

$$\sum_{j=1}^{L^p} \sum_{\{i \mid \Omega_j^{\#} \subset \tilde{\Omega}_i\}} \tilde{a}_{\sigma,\Omega_j^{\#}}(v|_{\Omega_j^{\#}}, v|_{\Omega_j^{\#}}) \leq \tilde{k}_0 \sum_{j=1}^{L^p} \tilde{a}_{\sigma,\Omega_j^{\#}}(v|_{\Omega_j^{\#}}, v|_{\Omega_j^{\#}}) = \tilde{k}_0 \|v\|^2_{a_\sigma}. \tag{4.39}$$

□

The following result is also essential but requires no change in the proof since it relies on the coercivity of a_σ on V_h, which is given in our case.

Lemma 4.3 (strengthened triangle inequality under the square, c.f. [36, Def. 3.6] [98, Lem. 7.9])**.** *For any collection of* $\{v_i\}_{i=1}^N$ *with* $v_i \in V_{h,0}(\tilde{\Omega}_i)$*, it holds that*

$$\left\| \sum_{i=1}^{N} v_i \right\|_{a_\sigma}^2 \leq N_\mathrm{c} \sum_{i=1}^{N} \|v_i\|_{a_\sigma}^2. \tag{4.40}$$

Proof. We follow the strategy from [98, Lem. 7.9] and expand the sum of Eq. (4.40) to remove all zero terms for domains with the same color according to the coloring map c from Eq. (4.25). Thus, let the set

$$\mathcal{C} := \left\{ (i,j) \in \{1,\dots,N\}^2 \;\middle|\; c(i) \neq c(j) \vee i = j \right\}. \tag{4.41}$$

Since a_σ is an inner product that induces a norm on V_h, we then apply the Cauchy–Schwarz inequality to obtain

$$\left\| \sum_{i=1}^{N} v_i \right\|_{a_\sigma}^2 = \sum_{(i,j)\in\mathcal{C}} a_\sigma(v_i, v_j) \leq \sum_{(i,j)\in\mathcal{C}} \|v_i\|_{a_\sigma} \|v_j\|_{a_\sigma}. \tag{4.42}$$

Defining the symmetric neighborhood matrix $\boldsymbol{C} \in \{0,1\}^{N\times N}$ using the indicator function as $C_{ij} = \chi_{\mathcal{C}}(i,j)$ and the vector $\boldsymbol{u} \in \mathbb{R}^N$ with $u_i = \|v_i\|_{a_\sigma}$, we can thus refine the relation Eq. (4.42) as

$$\left\| \sum_{i=1}^{N} v_i \right\|_{a_\sigma}^2 \leq \boldsymbol{u}^T \boldsymbol{C} \boldsymbol{u} \leq \lambda_{\max}(\boldsymbol{C}) \|\boldsymbol{u}\|_2^2 \leq \|\boldsymbol{C}\|_\infty \|\boldsymbol{u}\|_2^2 = \|\boldsymbol{C}\|_\infty \sum_{i=1}^{N} \|v_i\|_{a_\sigma}^2, \tag{4.43}$$

using Gershgorin's theorem. Then, the claim follows with $\|\boldsymbol{C}\|_\infty = N_\mathrm{c}$. □

The remaining part is the stability of the local contributions, for which we need the following result.

Lemma 4.4 (trivial kernel intersection)**.** *Assume (B3) and (B4), then for the NGEVPs from Definition 4.10,* $\ker \tilde{a}_{\sigma,\tilde{\Omega}_i} \cap \ker \tilde{b}_{\tilde{\Omega}_i} = \{0\}$ *for all* $i \in \tilde{\mathcal{I}}^\mathrm{i}$*.*

Proof. Let $i \in \tilde{\mathcal{I}}^\mathrm{i}$. By Lemma 4.1, we know that, on $V_h(\tilde{\Omega}_i)$, $\ker \tilde{a}_{\sigma,\tilde{\Omega}_i} = \operatorname{span}\{\tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi\}$. Since $\tilde{\chi}_i$ is never a constant due to the overlapping decomposition, we have $\tilde{\chi}_i \tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi \notin \operatorname{span}\{\tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi\}$, which implies that $\ker \tilde{a}_{\sigma,\tilde{\Omega}_i} \cap \ker \tilde{b}_{\tilde{\Omega}_i} = \{0\}$. □

The stable splitting for the coarse space from Definition 4.7 is achieved with the following projection.

Lemma 4.5 (local stability)**.** *Let $i \in \tilde{\mathcal{I}}^{\mathrm{i}}$ be given and assume (B3) and (B4), then the local projection operator $\tilde{\Pi}_{1,i} : V_h(\tilde{\Omega}_i) \to V_h(\tilde{\Omega}_i)$ with*

$$v \mapsto \tilde{\Pi}_{1,i} v := \frac{\tilde{b}_{\tilde{\Omega}_i}(v, \tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi)}{\tilde{b}_{\tilde{\Omega}_i}(\tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi, \tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi)} \tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi, \tag{4.44}$$

satisfies

$$\begin{aligned} |\tilde{\Pi}_{1,i} v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}} \leq |v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}}, \quad |v - \tilde{\Pi}_{1,i} v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}} \leq |v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}} \quad & \forall v \in V_h(\tilde{\Omega}_i), \\ |v - \tilde{\Pi}_{1,i} v|^2_{\tilde{b}_{\tilde{\Omega}_i}} \leq \frac{1}{\tilde{\lambda}_i^{(2)}} |v - \tilde{\Pi}_{1,i} v|^2_{\tilde{a}_{\sigma,\tilde{\Omega}_i}} \quad & \forall v \in V_h(\tilde{\Omega}_i), \end{aligned} \tag{4.45}$$

with $\tilde{\lambda}_i^{(2)} > 0$ from the Definition 4.10.

Proof. The proof follows by applying [36, Lem. 3.15], where we have the trivial kernel intersection satisfied by Lemma 4.4, the first unique eigenfunction given by $\tilde{p}_i^{(1)} = \tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi$ corresponding to a zero eigenvalue, and the positive semidefiniteness of $\tilde{a}_{\sigma,\tilde{\Omega}_i}$ and $\tilde{b}_{\tilde{\Omega}_i}$ on $V_h(\tilde{\Omega}_i)$. □

Note that the periodic coarse space only contains components for subdomains Ω_i, whose periodic neighborhood $\tilde{\Omega}_i$ does not touch the $\boldsymbol{x}$-boundary. No projection must be applied for the remaining subdomains since we have the following stability estimate.

Lemma 4.6. *Let $i \in \tilde{\mathcal{I}}^{\mathrm{b}}$, then the stability estimate $|v|^2_{\tilde{b}_{\tilde{\Omega}_i}} \leq \frac{1}{\tilde{\lambda}_i^{(1)}} |v|^2_{\tilde{a}_{\sigma,\tilde{\Omega}_i}}$ holds for all $v \in V_h(\tilde{\Omega}_i)$ with $\tilde{\lambda}_i^{(1)} > 0$ from the Definition 4.10.*

Proof. Let $i \in \tilde{\mathcal{I}}^{\mathrm{b}}$ be arbitrary. We apply [36, Lem. 3.15] using the projection on zero and $\tilde{\lambda}_i^{(1)} > 0$ since $V_h(\tilde{\Omega}_i)$ contains functions with zero boundary conditions in $\boldsymbol{x}$-direction, which can not be in $\ker \tilde{a}_{\sigma,\tilde{\Omega}_i} = \tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi$ with $\tilde{r}_i E_{\boldsymbol{x}}^{\#} \psi > 0$ on $\partial_{\boldsymbol{x}} \tilde{\Omega}_i$. □

The local stability of the projections Eq. (4.44) allows for a stable decomposition.

Theorem 4.1. *Assume (B3) and (B4), let $v \in V_h$, then the splitting defined by*

$$v_0 = \sum_{i \in \tilde{\mathcal{I}}^{\mathrm{i}}} \chi_i \tilde{\Pi}_{1,i} \tilde{r}_i v \in V_0, \quad v_i = \begin{cases} \chi_i (1 - \tilde{\Pi}_{1,i}) \tilde{r}_i v & i \in \tilde{\mathcal{I}}^{\mathrm{i}} \\ \chi_i \tilde{r}_i v & i \in \tilde{\mathcal{I}}^{\mathrm{b}} \end{cases} \in V_{h,0}(\Omega_i), \tag{4.46}$$

is C_0-stable with $C_0 = 2 + C_1 \tilde{k}_0 (2N_{\mathrm{c}} + 1)$, $C_1 := (\min \{\min_{i \in \tilde{\mathcal{I}}^{\mathrm{i}}} \tilde{\lambda}_i^{(2)}, \min_{i \in \tilde{\mathcal{I}}^{\mathrm{b}}} \tilde{\lambda}_i^{(1)}\})^{-1}$.

Proof. First, the splitting is consistent since, for all $v \in V_h$, we have $v_0 + \sum_{i=1}^N v_i = \sum_{i \in \tilde{\mathcal{I}}^{\mathrm{i}}} \chi_i \tilde{r}_i v + \sum_{i \in \tilde{\mathcal{I}}^{\mathrm{b}}} \chi_i \tilde{r}_i v = \sum_{i=1}^N \chi_i \tilde{r}_i v = \sum_{i=1}^N \chi_i r_i v = v$ using that $\tilde{r}_i = r_i$ on

$\mathrm{supp}(\chi_i)$. We have, using the strategy of [36, p13], that the splitting is locally stable since $\forall i \in \tilde{\mathcal{I}}^{\mathrm{i}}$, the local stability (Lemma 4.5), and the definition of $\tilde{b}_{\tilde{\Omega}_i}$ yields

$$\begin{aligned}\|v_i\|_{a_\sigma}^2 &= |\chi_i(1-\tilde{\Pi}_{1,i})\tilde{r}_i v|_{a_{\sigma,\Omega_i}}^2 = |\tilde{\chi}_i(1-\tilde{\Pi}_{1,i})\tilde{r}_i v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}}^2 \\ &= |\tilde{\chi}_i(1-\tilde{\Pi}_{1,i})\tilde{r}_i v|_{\tilde{a}_{0,\tilde{\Omega}_i}}^2 - \sigma\|\tilde{\chi}_i(1-\tilde{\Pi}_{1,i})\tilde{r}_i v\|_{L^2(\tilde{\Omega}_i)}^2 \leq |\tilde{\chi}_i(1-\tilde{\Pi}_{1,i})\tilde{r}_i v|_{\tilde{a}_{0,\tilde{\Omega}_i}}^2 \\ &= |(1-\tilde{\Pi}_{1,i})\tilde{r}_i v|_{\tilde{b}_{\tilde{\Omega}_i}}^2 \leq \frac{1}{\tilde{\lambda}_i^{(2)}}|(1-\tilde{\Pi}_{1,i})\tilde{r}_i v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}}^2 \leq \frac{1}{\tilde{\lambda}_i^{(2)}}|\tilde{r}_i v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}}^2.\end{aligned} \tag{4.47}$$

For all boundary domains with $i \in \tilde{\mathcal{I}}^{\mathrm{b}}$, Lemma 4.6 applies, and we get

$$\begin{aligned}\|v_i\|_{a_\sigma}^2 &= |\chi_i\tilde{r}_i v|_{a_{\sigma,\Omega_i}}^2 = |\tilde{\chi}_i\tilde{r}_i v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}}^2 = |\tilde{\chi}_i\tilde{r}_i v|_{\tilde{a}_{0,\tilde{\Omega}_i}}^2 - \sigma\|\tilde{\chi}_i\tilde{r}_i v\|_{L^2(\tilde{\Omega}_i)}^2 \\ &\leq |\tilde{\chi}_i\tilde{r}_i v|_{\tilde{a}_{0,\tilde{\Omega}_i}}^2 = |\tilde{r}_i v|_{\tilde{b}_{\tilde{\Omega}_i}}^2 \leq \frac{1}{\tilde{\lambda}_i^{(1)}}|\tilde{r}_i v|_{\tilde{a}_{\sigma,\tilde{\Omega}_i}}^2.\end{aligned} \tag{4.48}$$

Similar to [235, Lemma 2.9], using Eqs. (4.47) and (4.48), the SPSD splitting Eq. (4.37), and $\tilde{\mathcal{I}}^{\mathrm{i}} \cap \tilde{\mathcal{I}}^{\mathrm{b}} = \emptyset$ yields

$$\begin{aligned}\sum_{i=1}^N \|v_i\|_{a_\sigma}^2 &= \sum_{i\in\tilde{\mathcal{I}}^{\mathrm{i}}} \|v_i\|_{a_\sigma}^2 + \sum_{i\in\tilde{\mathcal{I}}^{\mathrm{b}}} \|v_i\|_{a_\sigma}^2 \leq \sum_{i\in\tilde{\mathcal{I}}^{\mathrm{i}}} \frac{1}{\tilde{\lambda}_i^{(2)}}|\tilde{r}_i v|_{a_{\sigma,\tilde{\Omega}_i}}^2 + \sum_{i\in\tilde{\mathcal{I}}^{\mathrm{b}}} \frac{1}{\tilde{\lambda}_i^{(1)}}|\tilde{r}_i v|_{a_{\sigma,\tilde{\Omega}_i}}^2 \\ \leq \frac{1}{\min_{i\in\tilde{\mathcal{I}}^{\mathrm{i}}}\tilde{\lambda}_i^{(2)}} \sum_{i\in\tilde{\mathcal{I}}^{\mathrm{i}}} |\tilde{r}_i v|_{a_{\sigma,\tilde{\Omega}_i}}^2 &+ \frac{1}{\min_{i\in\tilde{\mathcal{I}}^{\mathrm{b}}}\tilde{\lambda}_i^{(1)}} \sum_{i\in\tilde{\mathcal{I}}^{\mathrm{b}}} |\tilde{r}_i v|_{a_{\sigma,\tilde{\Omega}_i}}^2 \leq C_1 \sum_{i=1}^N |\tilde{r}_i v|_{a_{\sigma,\tilde{\Omega}_i}}^2 \leq C_1\tilde{k}_0\|v\|_{a_\sigma}^2,\end{aligned} \tag{4.49}$$

with $C_1 := (\min\{\min_{i\in\tilde{\mathcal{I}}^{\mathrm{i}}}\tilde{\lambda}_i^{(2)}, \min_{i\in\tilde{\mathcal{I}}^{\mathrm{b}}}\tilde{\lambda}_i^{(1)}\})^{-1}$. Then, we continue with

$$\begin{aligned}\|v_0\|_{a_\sigma}^2 &= \left\|v - \sum_{i=1}^N v_i\right\|_{a_\sigma}^2 \leq 2\|v\|_{a_\sigma}^2 + 2\left\|\sum_{i=1}^N v_i\right\|_{a_\sigma}^2 \leq 2\|v\|_{a_\sigma}^2 + 2N_{\mathrm{c}}\sum_{i=1}^N \|v_i\|_{a_\sigma}^2 \\ &\leq 2(1+N_{\mathrm{c}}C_1\tilde{k}_0)\|v\|_{a_\sigma}^2,\end{aligned} \tag{4.50}$$

using the relation Eq. (4.49) and the strengthened triangle inequality (Lemma 4.3). Combining the relations Eqs. (4.49) and (4.50) yields the C_0-stability as

$$\sum_{i=1}^N \|v_i\|_{a_\sigma}^2 + \|v_0\|_{a_\sigma}^2 \leq (2 + C_1\tilde{k}_0(2N_{\mathrm{c}}+1))\|v\|_{a_\sigma}^2. \tag{4.51}$$

□

With the C_0-stable decomposition, we finally obtain the following.

Theorem 4.2. *Assume (B3) and (B4), let V_0 be given by Definition 4.7, $\boldsymbol{M}_{\mathrm{AS},2}^{-1}$ by Eq. (4.24), N_{c} as in Definition 4.2, and C_0 as in Theorem 4.1. Then, the two-level additive Schwarz method satisfies the condition number bound*

$$\kappa(\boldsymbol{M}_{\mathrm{AS},2}^{-1}\boldsymbol{A}_\sigma) \leq C_0^2(N_{\mathrm{c}}+1). \tag{4.52}$$

Proof. The proof follows from [98, Cor. 5.12] and the Theorem 4.1. □

Remark 4.6. The condition number bound Eq. (4.52) becomes L-uniform for the natural decomposition using the L^p shifted unit-cells $\{\Omega_i^{\#}\}_{i=1}^N$ from Eq. (4.27), since N_c, $\tilde{k}_0$, and C_1 are L-invariant in that case.

4.4 Discussion About General Periodic Potentials

If the conditions (B3) and (B4) for the potential and the discretization are not fulfilled, some aspects of the analysis will change. For example, we can no longer use the fact that the first Neumann and the first periodic eigenfunction coincide. Moreover, the local Neumann problems may have negative eigenvalues. This has the following implications for the main ingredients of the theory.

- The proof of the SPSD splitting Eq. (4.37) for the periodic neighborhood uses the positive semidefiniteness on $V_h(\Omega_i)$. This might make it possible to locally reverse the factorization principle, leading to boundary terms that could cancel when using a proper alignment of the domain decomposition, similar to the periodic alignment in Definition 4.5.
- The strengthened triangle inequality Eq. (4.40) would still hold since the functions v_i are in $V_{h,0}(\Omega_i)$, on which the local bilinear forms are still positive definite since the global problem remains positive definite on V_h.
- The local stability of the projections in Lemma 4.5 would also need adaptation since its proof also requires positive definiteness. However, numerical results suggest that there might be only one negative Neumann eigenfunction of the shifted Schrödinger operator, whose eigenvalue is closer to zero than the second Neumann eigenvalue. Projecting on the zero-energy function ψ then would still yield local stability since the bilinear form would be elliptic on the orthogonal complement of the negative eigenspace.

Quantifying and exploring this behavior needs more investigation and remains future work. We note, however, that having global definiteness in combination with local indefiniteness is not a standard case to tackle. Other ideas to generalize the theory to the local indefinite case might be to use folding techniques, i.e., spectral coarse space approaches based on squaring the matrix, which represents fourth-order operators, to obtain a non-negative spectrum, which is inspired by, e.g. [41].

4.5 Numerical Experiments

In this section, we evaluate the performance of the proposed PerFact preconditioner from Definition 4.1 and provide numerical evidence for the theoretical results from Section 4.3. We choose the `Gridap` finite element framework [30] within the Julia [44] language due to its high-level interface to directly specify weak forms – similar to the

`FEniCS` framework [12, 250] in Python. In all the following tests, we use the CG and GMRES implementations of the `IterativeSolvers.jl`[1] package. For reproducibility, the source code and all metadata are available at [249].

4.5.1 Source Problem With the Two-Level Preconditioner

We first evaluate the performance of the coarse space for the solution of a shifted Schrödinger-type source problem on a two-dimensional rectangle $\Omega_L = (0, L) \times (0, 1)$ with a constant source term $f(x, y) = 1$: Find $u_h \in V_h \subset H_0^1(\Omega_L)$ such that

$$a_\sigma(u_h, v_h) = (f, v_h)_{L^2(\Omega_L)}, \quad \forall v_h \in V_h, \tag{4.53}$$

where a_σ is given by Eq. (4.23). The potential is $V(x, y) = 10^2(\sin^2(\pi x))^2(\sin^2(\pi y))^2$ fulfilling (B1) to (B4). The calculations use $\mathbb{Q}_1$ finite elements on a regular cartesian grid with mesh size $h = 1/10$. The shift $\sigma \approx 19.32644$ and the periodic solution ψ_h are obtained on the unit cell of the same mesh with periodic boundary conditions in the x-direction, according to Eq. (4.12).

Thus, this isolated setup of Eq. (4.53) simulates one application of the QOSI preconditioner $\boldsymbol{A}_\sigma^{-1}$ applied to the right-hand-side vector $\boldsymbol{f}$ generated by $f(x, y)$ and only tests the linear system solution rather than the complete eigenvalue algorithm. For each domain size L, we choose a non-overlapping, structured decomposition into $N = L$ subdomains, $\{\Omega_i'\}_{i=1}^L$ by using $\Omega_i' = (i - 1, i) \times (0, 1)$. Extending each domain by one layer of elements yields the overlapping $\{\Omega_i\}_{i=1}^L$, thus covering $(i - 1 - h, i + h) \times (0, 1)$ for $i = 2, \ldots, L - 1$ and the x-boundary domains accordingly. The two-level preconditioner $\boldsymbol{M}_{\mathrm{AS},2}^{-1}$ is then used within the CG method, which uses an initial guess of one and the relative residual condition $\mathtt{rTOL} = 10^{-8}$, where $\|\boldsymbol{r}_k\|_2 \leq \mathtt{rTOL}\|\boldsymbol{r}_0\|_2$ to determine the convergence.

4.5.1.1 Convergence Rate Comparison

The Fig. 4.4 compares the resulting relative residuals within the convergence history between the one-level and two-level PerFact preconditioner. For $\boldsymbol{M}_{\mathrm{AS},1}^{-1}$, the drastic increase in iteration numbers for $L \to \infty$ is visible, while the coarse space within $\boldsymbol{M}_{\mathrm{AS},2}^{-1}$ leads to a bounded convergence rate, which is independent of L, confirming our theory from Section 4.3.

4.5.1.2 Parameter Study

We also use this test case to investigate the dependency of the methods for changing mesh sizes $h \in \{1/10, 1/20, 1/30\}$ and overlap layer thickness $\delta \in \{1, 2, 3\}$ in units of elements. Again, we compare the CG method iteration numbers for the one-level and the PerFact two-level DD preconditioner using the same tolerance, starting vector, and problem setup. With the V_h-interpolation operator, I_h, we use the distance-based

[1] `https://github.com/JuliaLinearAlgebra/IterativeSolvers.jl`, version `v0.9.3`.

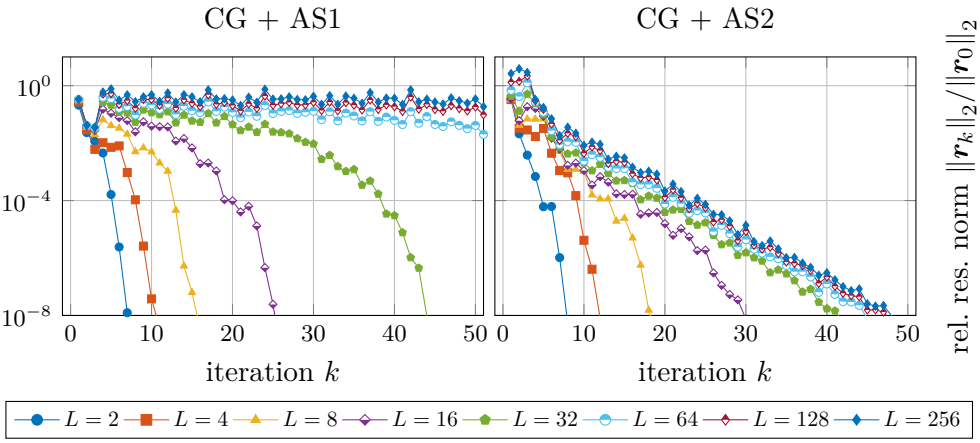


Figure 4.4: CG residual norms for varying domain length L using the AS preconditioner with no (left) and the PerFact coarse space (right).

(see, e.g., [98, Lem. 5.7]) partition of unity $\chi_i(\boldsymbol{z}) = I_h \circ (d_i(\boldsymbol{z})/\sum_{i=1}^N d_i(\boldsymbol{z}))$ with $d_i(\boldsymbol{z}) := \operatorname{dist}(\boldsymbol{z}, \partial\Omega_i \setminus \partial\Omega)\mathbf{1}_{\Omega_i}(\boldsymbol{z})$. In our case, the partition of unity resembles linear blending between two neighboring subdomains and allows keeping the gradients $\nabla\chi_i$ constant and thus minimal.

In Table 4.1, the theoretical results are confirmed for all δ and h since the PerFact coarse space is robust w.r.t. L. For moderate $L < 32$, both methods have comparable performance with the one-level preconditioner, even having some iterations less. This behavior is expected since the coarse space is applied additively and not multiplicatively (deflated) to resemble the case of the analysis. With the distance-based PU, increasing the overlap thickness δ is beneficial. Interestingly, keeping the ratio between subdomain size and overlap thickness $H/(\delta h)$ constant, i.e., $(h, \delta) \in \{(1/10, 1), (1/20, 2), (1/30, 3)\}$, does not change the convergence rate for a fixed L. Although this is a classical observation for, e.g., the Poisson problem, it is not directly clear for the quasi-optimally shifted Schrödinger operator since the shift σ also varies when h varies – which, in other words, means that we compare different operators in that case.

4.5.2 Linear Chain Model With Coulomb Potential

The initial hyperbox setup of Ω_L from Eq. (4.2) can be generalized to more complex domains $\hat{\Omega}$, which are generated by translated copies of a unit cell. Then, there exists an enclosing box such that $\hat{\Omega} \subset \Omega_L$ for some L. Thus, it is possible to consider a hypothetical penalty potential $\hat{V}(\boldsymbol{z}; V, a) := V(\boldsymbol{z}) + a\mathbf{1}_{\hat{\Omega}^c}(\boldsymbol{z})$ and take the limit of $a \to \infty$. Since the penalty is only applied in the complement $\hat{\Omega}^c := \Omega_L \setminus \hat{\Omega}$ when the indicator function $\mathbf{1}_{\hat{\Omega}^c}$ is nonzero, the resulting eigenfunctions of Eq. (4.1) converge to the eigenfunctions of the problem only solved in $\hat{\Omega}$ with zero Dirichlet conditions on $\partial\hat{\Omega}$ (see the *barrier principle* in Section 3.4.3.1).

Table 4.1: CG iterations for relative residuals to converge to $\texttt{rTOL} = 10^{-8}$ using no coarse space (first number) and the PerFact coarse space (second number).

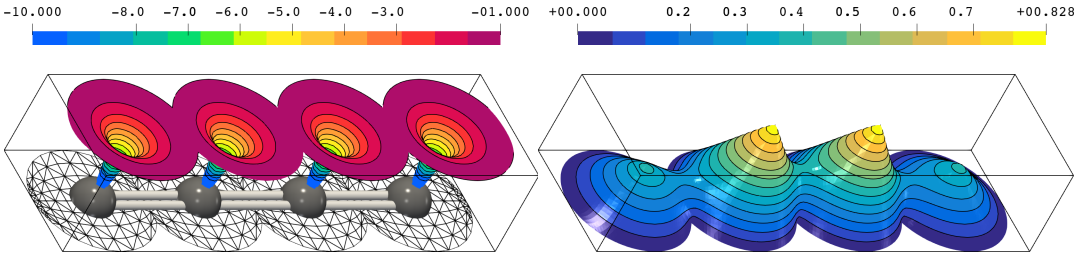
	$h = 1/10$			$h = 1/20$			$h = 1/30$		
L	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 1$	$\delta = 2$	$\delta = 3$	$\delta = 1$	$\delta = 2$	$\delta = 3$
2	5/6	4/5	4/5	6/7	6/6	5/6	8/8	6/7	6/7
4	7/9	6/8	6/8	9/11	8/9	7/9	11/12	9/11	9/10
8	11/14	10/13	10/12	14/17	12/15	11/15	16/19	14/17	13/15
16	19/22	17/19	15/17	23/26	20/23	19/21	26/30	23/25	21/23
32	32/28	28/23	25/21	40/36	35/31	32/27	44/42	40/35	37/30
64	56/31	48/25	43/21	70/40	61/33	56/27	80/45	70/38	63/31
128	100/33	88/25	79/22	129/42	111/34	100/28	149/48	129/38	115/32
256	188/33	164/25	149/22	241/42	208/34	187/28	285/48	241/38	217/34

4.5.2.1 Model Description

With that, we consider a linear chain of N particles, a toy model inspired by the cumulene configuration of the ideal carbyne [271]. Let $\{\boldsymbol{c}_i\}_{i=1}^N$ be a set of center positions with $\boldsymbol{c}_i := (R + 2(i-1)r, 0)^T$ for some $R > r > 0$, to construct a union of disks domain $\hat{\Omega}_N := \bigcup_{i=1}^N B_R(\boldsymbol{c}_i)$ with $B_R(\boldsymbol{c})$ being a disk of radius R and center $\boldsymbol{c} \in \mathbb{R}^2$. The domain can naturally be split into DD-suitable overlapping disks, following the spirit of [65]. Each particle generates a radial potential, modeled by a truncated Coulomb potential V_C, around its center fulfilling (B2). We find the unit cell to be $\hat{\Omega}_0 = (R - r, R + r) \times (-R, R)$ and note that it is not a full circle, leading to defect regions in the mesh at both x-ends. However, the factorization theory (see the *principle of defect invariance* in Section 3.4.3.2) and the DD analysis from Section 4.3 still apply. The neighboring interactions between two particles are neglected for simplicity, but we must add the two boundary ghost centers to the collection $\boldsymbol{\mathcal{C}} = \{\boldsymbol{c}_i\}_{i=1}^N \cup \{(R-2r, 0)^T, (R+2Nr, 0)^T\}$ to fulfill the assumption (B1). The resulting potential then reads $V(\boldsymbol{z}) = V_\mathrm{C}\left(\min_{\boldsymbol{c}\in\boldsymbol{\mathcal{C}}} \|\boldsymbol{z} - \boldsymbol{c}\|_2\right)$ with $V_\mathrm{C}(r) = -Z/\max\{r, b\}$ for some parameters $Z, b > 0$. A visualization of the potential and an exemplary $\mathbb{P}_1$ finite element mesh is given in Fig. 4.5a.

4.5.2.2 Flexibility Test of the Coarse Space Within Eigensolvers

With the union of disks setup, we now test the coarse space within the SI-LOPCG solver from Algorithm 2. To demonstrate the flexibility of the PerFact coarse space, we not only test the ASM2 method but also consider the stationary RAS2 iteration (multiplicative coarse correction) from Eq. (4.18) as well as the RAS2 preconditioner from Eqs. (4.15) and (4.17) with the PerFact coarse space from Eq. (4.20) to use within the GMRES method (restart not reached). For the parameters $Z = 1, b = 10^{-4}$, a series of computations is then performed for $N \in \{1, 2, 4, \ldots, 128\}$ on meshes with x-length proportional to N while the coarse space dimension is $|V_0| = N$. The discretization results in 1353 nodes for the unit cell $\hat{\Omega}_0$, and the resulting solution



(a) Potential function and union of disks mesh.

(b) First eigenfunction ϕ for $N = 4$.

Figure 4.5: A union of disks domain Ω_N for $N = 4$ with **(a)** the applied symmetric potential V and an exemplary $\mathbb{P}_1$ mesh and **(b)** the resulting first eigenfunction ϕ. Both color scales divided the listed interval into 14 colors.

is focused in the center of the domain. A visualization of the first eigenfunction for $N = 4$ is given in Fig. 4.5b.

The resulting iteration numbers are reported in Table 4.2. First, we observe that the one-level RAS1 is, as expected, not robust for increasing L, such that the computations are skipped for $L > 8$. The stationary RAS2 method performs reasonably well and can be used as an inner solver. However, the Krylov accelerated CG+ASM2 and GMRES+RAS2 variants work better, having much fewer inner iterations per outer step. They are also robust with respect to L since the maximum number of inner iterations, denoted by $\max_\mathrm{i}$, is bounded from above. Combined with bounded outer iterations obtained by the QOSI strategy, this yields a bounded number of inner iterations, denoted by $\sum_\mathrm{i}$, which measures computational cost. We also observe that the unsymmetric GMRES+RAS2 case has fewer iterations than the symmetric CG+ASM2 case – which is the expected behavior. Furthermore, the outer iterations are not influenced by the inner solver. Another interesting observation relates to cases when the maximum number of inner iterations $k_\mathrm{max} = 1000$ is reached for the RAS1 ($N = 8$) or the RAS2 ($L = 128$) method. Here, we observe that even in these cases, convergence of the eigenvalue solver can still be achieved. This opens questions about the interplay between inner and outer tolerances, which we discuss in the following.

4.5.3 Fusing the Loops

The inner tolerance can be made adaptive to further decrease the total number of inner iterations. Instead of the fixed tolerance of $\texttt{rTOL}_\mathrm{i} = 10^{-8}$, we choose it to be proportional to the outer residual norm as $\texttt{rTOL}_\mathrm{i} = \min\{0.1, \|\boldsymbol{r}_{\mathrm{o},k}\|_2\}$, inspired by [114]. We compare these two cases with a third approach, which skips the inner Krylov solver and directly uses $\boldsymbol{M}_{\mathrm{RAS},2}^{-1}$ as a preconditioner in the LOPCG method, i.e., solving $\boldsymbol{M}_{\mathrm{RAS},2}\boldsymbol{w}_k = \boldsymbol{r}_{\mathrm{o},k}$ in line 4 of Algorithm 2 (instead of $\boldsymbol{A}_\sigma\boldsymbol{w}_k = \boldsymbol{r}_{\mathrm{o},k}$). This strategy still has the shift-and-invert effect and does not suffer from a high condition number.

To render the case more complex, a three-dimensional plane-like box $\Omega_L = (0, L)^2 \times (0, 1)$ is considered with a potential V that is periodic in the x- and y-directions and

Table 4.2: Inner and outer iteration numbers of the SI-LOPCG method using the stationary RAS1, stationary RAS2, CG+ASM2, and GMRES+RAS2 as inner solvers. Skipped simulations are indicated with †. We apply a relative tolerance of $\texttt{rTOL}_\text{i} = 10^{-10}$ for the inner residuals and an absolute tolerance (since the eigenvector is normalized after each iteration) of $\texttt{TOL}_\text{o} = 10^{-8}$ for the outer spectral residuals. For the inner solver, $k_\text{max} = 1000$ applies, and 1000* is displayed when k_max is reached. We abbreviate with it_o the outer iterations, with $\max_\text{i}$ the maximal number of inner iterations as a measure for the worst case, and with $\sum_\text{i}$ the sum of all inner iterations (approximately computational costs).

	RAS1			RAS2			CG+ASM2			GMRES+RAS2		
$N \sim L$	it_o	$\max_\text{i}$	$\sum_\text{i}$	it_o	$\max_\text{i}$	$\sum_\text{i}$	it_o	$\max_\text{i}$	$\sum_\text{i}$	it_o	$\max_\text{i}$	$\sum_\text{i}$
1	5	38	174	5	21	96	5	17	82	5	14	68
2	5	195	865	5	117	519	5	20	98	5	17	83
4	4	650	2278	4	157	557	4	27	103	4	22	85
8	5	1000*	5000	5	180	765	5	39	186	5	33	155
16	†	†	†	5	192	818	5	51	251	5	46	213
32	†	†	†	4	200	732	4	65	260	4	59	229
64	†	†	†	4	206	766	4	89	342	4	70	272
128	†	†	†	4	1000*	1585	4	90	351	4	70	267

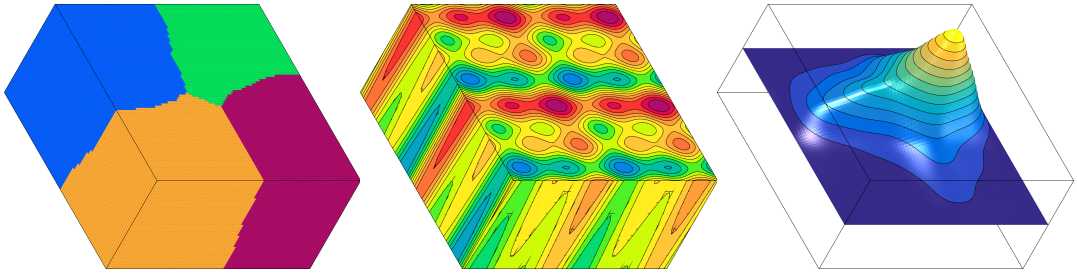
has a linear gradient in the z-direction. On purpose, we now choose an unsymmetric potential (not fulfilling (B4)), which is given by

$$V(x, y, z) = 10(4 + \sin 2\pi x + \sin 4\pi x + 2\sin 2\pi y + 2\sin 4\pi y + z). \tag{4.54}$$

For the domain decomposition, we also apply the most general case by using an unstructured METIS partition. The Fig. 4.6 presents the partition, V, and an exemplary first eigenfunction for the case of $L = 2$. With a structured $\mathbb{Q}_1$-discretization using $h = 1/10, \delta = 1, \texttt{TOL}_\text{o} = 10^{-10}$ and inner initial guess of zero, we perform a series of computations for $L \in \{4, 8, 16, 32\}$ and keep track of all inner iterations. As we observe in Fig. 4.7, the method is still robust w.r.t. to L, even for the unsymmetric potential case. Although the adaptive strategy can significantly reduce the number of iterations, the fused approach leads to the fastest convergence. The idea of fusing the inner and the outer loop could even be carried further to the case of nonlinear eigenvalue problems, which have an additional third loop to handle the nonlinearity.

4.6 Conclusion and Future Work

In this chapter, we presented a new domain decomposition preconditioner for the quasi-optimally shifted Schrödinger operator, which is robust with respect to anisotropic domain expansion. This setup is motivated by 1d structures (e.g., carbon nanotubes) or 2d materials (e.g., graphene) in material science. Initially, we academically motivated



(a) METIS partition of Ω_2. (b) Potential V (unsymmetric). (c) First eigenfunction ($L = 2$).

Figure 4.6: **(a)** An unstructured METIS element partition of Ω_2 into $\{\Omega_i'\}_{i=1}^4$, **(b)** a periodic but unsymmetric potential V, and **(c)** the first eigenfunction for the case of $L = 2$ with $h = 1/40$.

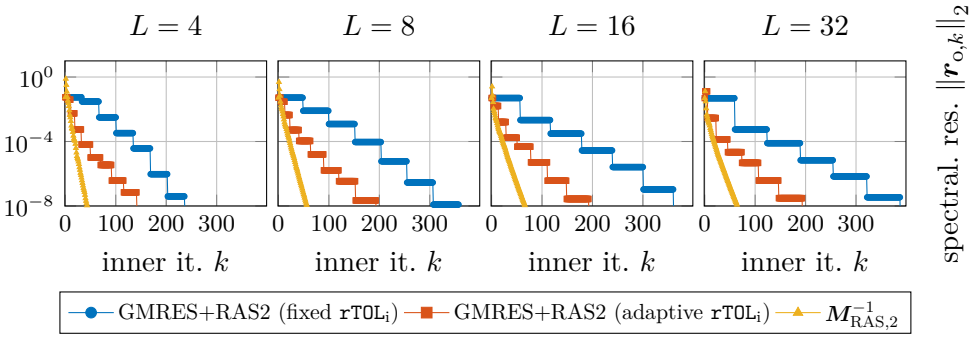


Figure 4.7: Comparison of the total number of inner iterations for a fixed inner tolerance, an adaptive inner tolerance, and direct usage of the RAS2-preconditioner within the LOPCG method.

the fundamental dilemma of effective shifting for eigenvalue problems in contrast to the fast convergence of iterative eigenvalue solvers. With the new factorization preconditioner, we successfully combined both aspects to simultaneously achieve fast convergence of both the eigenvalue and linear solvers. For our analysis, we utilized tools from the theory of spectral coarse spaces to prove a condition number bound for the preconditioned system. Operating in a setup where coercivity asymptotically vanishes, we took special care in adapting the geometrical setup and the theory accordingly, for instance, by considering a hypothetical alignment of the domain decomposition. The theoretical results are confirmed by numerical experiments, which also demonstrate the flexibility of the coarse space within eigensolvers.

Limitations of the method include the assumptions on the potentials required to apply the theory and the periodicity assumption. Consequently, future work can focus on extending the theory to more general potentials and applying it to more realistic nonlinear eigenvalue problems, such as tight-binding models. Furthermore, rather than solving only one problem on a unit cell to obtain a shift and coarse space, we

could consider general parameter-dependent eigenvalue problems. In such a scenario, known coarse spaces from the solution of linear systems might serve as an efficient basis to extract a suitable shift for use within the eigenvalue solver. In other words, future efforts could strengthen the connection between solver theory from iterative linear systems to eigenvalue solvers.

5 Conclusion, Outlook, and Future Work

In this thesis, we presented strategies to solve the linear Schrödinger eigenvalue problem for anisotropic structures with a vanishing fundamental gap. Due to the very diverse use and combination of techniques, let us recall the main contributions of this thesis and discuss weaknesses, future research directions, and open questions that arise when considering the specific problems within this thesis and go beyond them when stepping back and looking at the general picture.

5.1 Summary

In Chapter 1, we motivated the general need to construct scalable algorithms and provided a broad context of what motivated this work. We also highlighted the main contributions of this thesis and provided an overview of how these built up on each other.

In Chapter 2, we first provided the physical background of the linear Schrödinger equation model. We started with the comprehensive picture of the quantum system description and then moved on to modeling many-body molecular systems. Due to their complexity, we reviewed and explained the relevant model assumptions and approximations to obtain the electronic ground state problem. Particular focus was then given to the presentation of the Hartree–Fock and the Kohn–Sham models, leading to a system of nonlinear eigenvalue problems. This presentation allowed us to embed the mathematical model problem of the linear Schrödinger equation, used within the later chapters, into the broader context. The role of anisotropic domains was discussed from a practical viewpoint, and difficulties were shown. We then moved on to the discretization of the equations with the subsequent presentation of iterative algorithms to solve the resulting discrete eigenvalue problems. We mainly focussed on single vector iterations and gradient-based algorithms. Finally, preconditioning was discussed, and the domain decomposition method was introduced.

Having all these tools at hand, we started in Chapter 3 with the first main contribution of this thesis. We presented a quasi-optimal shift-invert preconditioner for the linear Schrödinger equation with a periodic potential. We stated all model assumptions and used the simple Laplace eigenvalue problem as an illustrative example to highlight the significant difficulties. After outlining the chapter-specific context, we defined a prototype problem and derived a factorization result for the eigenpairs,

which is especially useful for extracting information about their asymptotic behavior. The factorization was also illustrated graphically to provide a better understanding. Afterward, this result was used to factorize the problem in a size-independent part and another eigenvalue problem. For the remaining eigenvalue problem, we analyzed its asymptotic behavior after applying a rescaling. This analysis required the framework of directional homogenization theory since the rescaling leads to highly oscillating coefficients. Based on that analysis, we could show that the eigenvalue goes to zero, which implies that the first eigenvalue of our model problem has a limit, given by the solution to a cell problem. For the numerics, we constructed a quasi-optimal shift-and-invert (QOSI) preconditioner. As extensive numerical tests showed, this preconditioner yielded robustness in iteration numbers for the inverse power and the LOCG method. We also discussed the method's limitations, especially the ill-conditioning issue of the resulting shifted linear systems.

The work in the Chapter 4 is a direct continuation of the previous chapter in that it tackles the remaining problems to allow the use of iterative linear solvers. We again presented the model and all assumptions and used the Laplace eigenvalue problem to motivate and show the difficulties that arise. However, this time, we focussed on using inner-outer type eigenvalue solvers and measured the complexity in terms of the total number of inner iterations. We showed a fundamental shifting dilemma: although shifting speeds up the iterative eigenvalue solvers, it also increases the condition number of the resulting matrices. This behavior is not an artifact, but it is by definition. We then presented the chapter-specific context and moved on to the introduction to domain decomposition and how we use it in combination with an eigenvalue solver. Specifically, we treat the DD method as an inner solver or preconditioner for Krylov solvers. Since the one-level DD approach can not be used directly, we introduced a particular coarse space (PerFact) that used the same asymptotic analysis from the Chapter 3 using the limit eigenfunctions as generators. These coarse space components are already available and intuitively treat the lower, problematic part of the shifted spectrum. We then presented an analysis of the resulting two-level domain decomposition preconditioner under some symmetry assumptions. We observed that our coarse space is closely related to spectral coarse spaces and that we had to adapt our problem to the general framework. After all, we obtained a condition number bound independent of the domain size. Having this at hand, we implemented and tested the resulting method and confirmed the theory. The numerical tests showed that the method works for the general case with fewer assumptions. Finally, we presented a fusing strategy that replaces the two loops in the inner-outer strategy with a single loop. We tested the proposed preconditioner and obtained a significant reduction in the total number of inner iterations.

5.2 Outlook and Future Research Directions

Regarding future research directions, we can think of several things. Some of them are obvious and specifically related to the methods within this thesis. However, some are

also more general and ask more fundamental questions, especially about the role of preconditioning for eigenvalue algorithms. Let us start with the first set of questions.

5.2.1 Analysis of the Methods for More General Problems

We start with the most obvious question: How do the methods perform for more general problems? In this thesis, we have focused on the linear Schrödinger equation with periodic potentials. We can think of the following generalizations:

- **Beyond periodicity**: Since the potential V in Chapters 3 and 4 was assumed to be directional-periodic and thus neglects the interaction beyond the unit cells, we could formulate the problem with a more general potential as the sum of local potentials, each located at the lattice sites. This formulation would lead to a more general problem in a non-periodic setting. A potential of the form $V_L(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{L} V_\mathrm{c}(\boldsymbol{x} - \boldsymbol{c}_i, \boldsymbol{y})$ for some cell potential V_c and cell centers $\boldsymbol{c}_i$ would be a natural choice. First tests suggest that if the cell potential V_c is regular enough, e.g., by using a regularized Yukawa-type potential $V_\mathrm{c}(\boldsymbol{r}) = e^{-\alpha r}/(r + \beta)$ for some $\alpha, \beta > 0$ (in cell-centered radial coordinates) allows to take the limit of V_L for $L \to \infty$. The splitting of $V_L = V_\infty + \tilde{V}_L$ would then allow the factorization approach from Section 3.2 to be applied to the periodic part V_∞. At the same time, the non-periodic part $\tilde{V}_L$ would be treated as a perturbation that converges to zero almost everywhere.

- **Algebraic factorization**: The factorization approach from Section 3.2 was done in the infinite-dimensional case. A discrete analog of the factorization approach must also exist and could be analyzed. For these discrete cases, there might be further applications beyond the PDE setting, where specific periodicity effects play a role and need to be factorized to create a reduced model, e.g., when modeling large graphs or networks with a proper notation of periodicity. Considering the discrete matrix case might be closely related to tight-binding models, see [73, 160], where usually only nearest-neighbor interactions are considered, and the discretized Hamiltonian matrix is thus very structured.

- **Initial guess optimization**: In this thesis, we have optimized only the convergence rate of algorithms to make them robust for any starting value. However, we should also optimize the initial guesses since worse convergence rates can be compensated somewhat by better initial guesses. For the Laplace problem, we know the first eigenfunction for the abstract box geometry and can use it as a starting guess; by definition, the method is immediately converged. By accident, we found cases where using the first eigenfunction of the Laplace eigenvalue problem also for periodic Schrödinger operators was an excellent initial guess with orthogonality to the actual first excited Schrödinger eigenfunctions up to, at least, the index L (if $L \in \mathbb{N}$, for the two-dimensional case), i.e., $\langle \boldsymbol{u}^{(1)}_{-\Delta}, \boldsymbol{u}^{(i)}_{-\Delta+V} \rangle = 0$ for $i = 2, \ldots, L$. Since this only holds for specific symmetric V, an analysis could be based on the theory of *nodal lines* [131], i.e. the places

where an eigenfunction is zero, and how they divide the domain into equally sized parts, for instance. This analysis would be an exciting direction to explore and might lead, in combination with the DD approach, to a scalable method that would not require a coarse space.

- **Nonlinear problems**: The extension to the nonlinear case is a natural question. The primary problem is that the factorization approach only applies to linear problems. However, we could still use the external potential or extract a periodic part from it and then use the factorization approach from Section 3.2 on the linearized problem. Suppose the remaining contribution of the potential is positive. In that case, the resulting analysis will yield a lower bound on the first eigenvalue (at least for the GPE), which might be used as a shift in SCF iterations. However, asymptotically sharp bounds for the shifted fundamental ratio seem out of reach. Still, this would be an exciting approach to try out.

5.2.2 General Questions About Iterative Eigenvalue Solvers

Moving a step back and looking abstractly at the work of this thesis, we designed a method for quasi-optimal preconditioning of an eigenvalue problem with a vanishing fundamental gap. We found an interesting connection between the shifting parameter and the required coarse space components. So, we can ask more fundamental questions:

- **Connecting linear and eigensolver preconditioning**: For linear solvers, the performance of a preconditioner is measured in terms of spectral closeness to the inverse of the matrix to get bounds of the type $\kappa(\boldsymbol{M}^{-1}\boldsymbol{A}) < C$. Such a preconditioner is insufficient for gradient-based eigenvalue solvers since the spectral gap might vanish. Typical convergence results, see, e.g. [25], thus still include the spectral gap or the fundamental ratio between the first eigenvalues. Since the perfect preconditioner is $(\boldsymbol{A} - \lambda^{(1)}\boldsymbol{I})^{\dagger}$ (see Section 2.2.2), one could start the analysis from, e.g., the spectral equivalence $(1-\gamma)\boldsymbol{x}^T\boldsymbol{P}^{-1}\boldsymbol{x} \leq \boldsymbol{x}^T(\boldsymbol{A}-\lambda^{(1)}\boldsymbol{I})\boldsymbol{x} \leq (1+\gamma)\boldsymbol{x}^T\boldsymbol{P}^{-1}\boldsymbol{x}$ for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \operatorname{span}(\boldsymbol{x}^{(1)})$ or the requirement $\kappa((\boldsymbol{M}^{-1}(\boldsymbol{A}-\lambda^{(1)}\boldsymbol{I}))|_{E(\boldsymbol{x}^{(1)})^{\perp}}) < C$, with $E(\boldsymbol{x}^{(1)})^{\perp}$ denoting the orthogonal complement of the first eigenspace, in connection with the case of linear systems. Then, one could consider eigenvalue problems that depend on a general parameter p. This parameter could be related to, e.g., the domain size L, the mesh size h, the strength of the nonlinearity β (for GPE), the overlap parameter δ, the number of subdomains N_{sd}, a (random) diffusion coefficient $A(\boldsymbol{x})$. Then, a correspnding eigenvalue problem with p-dependency, $\boldsymbol{A}_p\boldsymbol{x}_p = \lambda_p\boldsymbol{x}_p$, could be considered. Asymptotic analysis for the lowest eigenvalue $\lambda_p^{(1)}$ as $p \to 0$ can then be used to derive an asymptotic limit for such a general eigenvalue problem – similar to what we did in this thesis for $p = 1/L$ by deriving the asymptotic limit of $\lambda_L^{(1)}$ as $L \to \infty$.

- **Using an approximate shift from the coarse space**: In Chapters 3 and 4, we derived a quasi-optimal shift and later used related coarse spaces to fix the

ill-conditioning issues for the shifted systems. However, the other way around is also possible. Various methods exist to construct coarse space, multiscale basis functions, or reduced models that all try to capture the extremal ends of the operator spectrum. In a Rayleigh–Ritz procedure, one could use these reduced models to extract an approximate shift, $\tilde{\sigma}$, as, e.g., the lowest eigenvalue in the coarse basis and then use it as a shift for the refined model. However, special care needs to be taken since, by the min-max theorem, $\lambda^{(1)} \leq \tilde{\sigma}$ as the coarse space is a subspace of the refined space. An analysis to provide guaranteed distance bounds to the real eigenvalue of interest might be another interesting direction to explore.

- **Convergence analysis of the fused iteration loop**: Lastly, we found out in the Section 4.5.3 that fusing the inner-outer loops within the eigensolver benefited the total number of inner linear system iterations. Analyzing this behavior might be another interesting direction to explore. It is closely related to the first question since the fused preconditioning strategy might represent an approximate preconditioner to the $\lambda^{(1)}$-shifted system.

Bibliography

[1] P. M. Ajayan and O. Z. Zhou. "Applications of Carbon Nanotubes". In: *Carbon Nanotubes: Synthesis, Structure, Properties, and Applications*. Ed. by M. S. Dresselhaus, G. Dresselhaus, and P. Avouris. Topics in Applied Physics. Berlin, Heidelberg: Springer, 2001, pp. 391–425. DOI: 10.1007/3-540-39947-X_14 (cit. on p. 42).

[2] H. Al Daas and P. Jolivet. "A Robust Algebraic Multilevel Domain Decomposition Preconditioner for Sparse Symmetric Positive Definite Matrices". In: *SIAM J. Sci. Comput.* 44.4 (2022), A2582–A2598. DOI: 10.1137/21M1446320 (cit. on pp. 37, 77).

[3] H. Al Daas, P. Jolivet, and T. Rees. "Efficient Algebraic Two-Level Schwarz Preconditioner for Sparse Matrices". In: *SIAM J. Sci. Comput.* 45.3 (2023), A1199–A1213. DOI: 10.1137/22M1469833 (cit. on pp. 36, 37, 77).

[4] G. Allaire. "A Brief Introduction to Homogenization and Miscellaneous Applications". In: *ESAIM: Proc.* 37 (2012), pp. 1–49. DOI: 10.1051/proc/201237001 (cit. on pp. 52, 53, 55, 56).

[5] G. Allaire. *Numerical Analysis and Optimization: An Introduction to Mathematical Modelling and Numerical Simulation*. Numerical Mathematics and Scientific Computation. Oxford University Press, 2007 (cit. on pp. 55, 56).

[6] G. Allaire and G. Bal. "Homogenization of the Criticality Spectral Equation in Neutron Transport". In: *ESAIM: M2AN* 33.4 (1999), pp. 721–746. DOI: 10.1051/m2an:1999160 (cit. on p. 42).

[7] G. Allaire and Y. Capdeboscq. "Homogenization and Localization for a 1-D Eigenvalue Problem in a Periodic Medium with an Interface". In: *Ann. Mat. Pura Appl. IV. Ser.* 181.3 (2002), pp. 247–282. DOI: 10.1007/s102310100040 (cit. on p. 42).

[8] G. Allaire and Y. Capdeboscq. "Homogenization of a Spectral Problem in Neutronic Multigroup Diffusion". In: *Comput. Methods Appl. Mech. Engrg.* 187.1 (2000), pp. 91–117. DOI: 10.1016/S0045-7825(99)00112-7 (cit. on pp. 42, 46, 56, 73).

[9] G. Allaire and C. Conca. "Bloch Wave Homogenization and Spectral Asymptotic Analysis". In: *Journal de Mathématiques Pures et Appliquées* 77.2 (1998), pp. 153–208. DOI: 10.1016/S0021-7824(98)80068-8 (cit. on p. 42).

[10] G. Allaire and F. Malige. "Analyse asymptotique spectrale d'un problème de diffusion neutronique". In: *Comptes Rendus de l'Académie des Sciences - Series I - Mathematics* 324.8 (1997), pp. 939–944. DOI: 10.1016/S0764-4442(97)86972-8 (cit. on pp. 42, 56).

[11] G. Allaire and A. Piatnitski. "Homogenization of the Schrödinger Equation and Effective Mass Theorems". In: *Commun. Math. Phys.* 258.1 (2005), pp. 1–22. DOI: 10.1007/s00220-005-1329-2 (cit. on p. 42).

[12] M. Alnæs, J. Blechta, J. Hake, A. Johansson, B. Kehlet, A. Logg, C. Richardson, J. Ring, M. E. Rognes, and G. N. Wells. "The FEniCS Project Version 1.5". In: *ans* 3.100 (2015). DOI: 10.11588/ans.2015.100.20553 (cit. on pp. 22, 60, 94).

[13] F. Alouges and C. Audouze. "Preconditioned Gradient Flows for Nonlinear Eigenvalue Problems and Application to the Hartree-Fock Functional". In: *Numerical Methods for Partial Differential Equations* 25.2 (2009), pp. 380–400. DOI: 10.1002/num.20347 (cit. on p. 23).

[14] R. Altmann and D. Peterseim. "Localized Computation of Eigenstates of Random Schrödinger Operators". In: *SIAM J. Sci. Comput.* 41.6 (2019), B1211–B1227. DOI: 10.1137/19M1252594 (cit. on p. 42).

[15] R. Altmann, D. Peterseim, and T. Stykel. *Riemannian Newton Methods for Energy Minimization Problems of Kohn-Sham Type*. 2023. DOI: 10.48550/arXiv.2307.13820. arXiv: 2307.13820 [cs, math] (cit. on pp. 23, 76).

[16] R. Altmann, P. Henning, and D. Peterseim. "Localization and Delocalization of Ground States of Bose–Einstein Condensates Under Disorder". In: *SIAM J. Appl. Math.* 82.1 (2022), pp. 330–358. DOI: 10.1137/20M1342434 (cit. on p. 42).

[17] R. Altmann, P. Henning, and D. Peterseim. "Quantitative Anderson Localization of Schrödinger Eigenstates under Disorder Potentials". In: *Math. Models Methods Appl. Sci.* 30.05 (2020), pp. 917–955. DOI: 10.1142/S0218202520500190 (cit. on p. 42).

[18] R. Altmann, P. Henning, and D. Peterseim. "The J-method for the Gross-Pitaevskii Eigenvalue Problem". In: *Numer. Math.* 148.3 (2021), pp. 575–610. DOI: 10.1007/s00211-021-01216-5 (cit. on pp. 42, 76).

[19] R. Altmann, D. Peterseim, and T. Stykel. "Energy-Adaptive Riemannian Optimization on the Stiefel Manifold". In: *ESAIM: M2AN* 56.5 (2022), pp. 1629–1653. DOI: 10.1051/m2an/2022036 (cit. on pp. 23, 76).

[20] X. Andrade, D. Strubbe, U. D. Giovannini, A. H. Larsen, M. J. T. Oliveira, J. Alberdi-Rodriguez, A. Varas, I. Theophilou, N. Helbig, M. J. Verstraete, L. Stella, F. Nogueira, A. Aspuru-Guzik, A. Castro, M. A. L. Marques, and A. Rubio. "Real-Space Grids and the Octopus Code as Tools for the Development of New Simulation Approaches for Electronic Systems". In: *Phys. Chem. Chem.*

Phys. 17.47 (2015), pp. 31371–31396. DOI: 10.1039/C5CP00351B (cit. on p. 20).

[21] X. Antoine and R. Duboscq. "GPELab, a Matlab Toolbox to Solve Gross-Pitaevskii Equations I: Computation of Stationary Solutions". In: *Computer Physics Communications* 185.11 (2014), pp. 2969–2991. DOI: 10.1016/j.cpc.2014.06.026 (cit. on p. 42).

[22] X. Antoine and R. Duboscq. "GPELab, a Matlab Toolbox to Solve Gross-Pitaevskii Equations II: Dynamics and Stochastic Simulations". In: *Computer Physics Communications* 193 (2015), pp. 95–117. DOI: 10.1016/j.cpc.2015.03.012 (cit. on p. 42).

[23] X. Antoine, A. Levitt, and Q. Tang. "Efficient Spectral Computation of the Stationary States of Rotating Bose-Einstein Condensates by the Preconditioned Nonlinear Conjugate Gradient Method". In: *Journal of Computational Physics* 343 (2017), pp. 92–109. DOI: 10.1016/j.jcp.2017.04.040 (cit. on p. 42).

[24] P. Arbenz. *Lecture Notes on Solving Large Scale Eigenvalue Problems (Spring 16)*. ETH Zürich. 2016 (cit. on pp. 31, 32).

[25] M. E. Argentati, A. V. Knyazev, K. Neymeyr, E. E. Ovtchinnikov, and M. Zhou. "Convergence Theory for Preconditioned Eigenvalue Solvers in a Nutshell". In: *Found. Comput. Math.* 17.3 (2017), pp. 713–727. DOI: 10.1007/s10208-015-9297-1. arXiv: 1412.5005 (cit. on pp. 30, 104).

[26] D. Arndt, W. Bangerth, D. Davydov, T. Heister, L. Heltai, M. Kronbichler, M. Maier, J.-P. Pelteret, B. Turcksin, and D. Wells. "The Deal.II Finite Element Library: Design, Features, and Insights". In: *Computers & Mathematics with Applications*. Development and Application of Open-source Software for Problems with Numerical PDEs 81 (2021), pp. 407–422. DOI: 10.1016/j.camwa.2020.02.022 (cit. on p. 22).

[27] A. P. Austin and L. N. Trefethen. "Computing Eigenvalues of Real Symmetric Matrices with Rational Filters in Real Arithmetic". In: *SIAM J. Sci. Comput.* 37.3 (2015), A1365–A1387. DOI: 10.1137/140984129 (cit. on p. 28).

[28] I. Babuška and J. E. Osborn. "Finite Element-Galerkin Approximation of the Eigenvalues and Eigenvectors of Selfadjoint Problems". In: *Mathematics of Computation* 52.186 (1989), p. 24. DOI: 10.1090/S0025-5718-1989-0962210-8 (cit. on p. 59).

[29] M. Bachmayr. "Low-Rank Tensor Methods for Partial Differential Equations". In: *Acta Numerica* 32 (2023), pp. 1–121. DOI: 10.1017/S0962492922000125 (cit. on pp. 19, 20).

[30] S. Badia and F. Verdugo. "Gridap: An Extensible Finite Element Toolbox in Julia". In: *JOSS* 5.52 (2020), p. 2520. DOI: 10.21105/joss.02520 (cit. on pp. 22, 60, 93).

[31] Z. Bai, J. Demmel, J. Dongarra, A. Ruhe, and H. van der Vorst, eds. *Templates for the Solution of Algebraic Eigenvalue Problems: A Practical Guide*. Society for Industrial and Applied Mathematics, 2000. DOI: 10.1137/1.9780898719581 (cit. on pp. 26, 28–30, 40, 59, 78, 79).

[32] W. Bao. "Mathematical Models and Numerical Methods for Bose-Einstein Condensation". In: *Proceedings of the International Congress of Mathematicians Seoul* (2014) (cit. on p. 19).

[33] W. Bao and Y. Cai. "Mathematical Theory and Numerical Methods for Bose-Einstein Condensation". In: *Kinetic & Related Models* 6.1 (2013), pp. 1–135. DOI: 10.3934/krm.2013.6.1 (cit. on p. 19).

[34] M. Barrault, G. Bencteux, E. Cancès, W. W. Hager, and C. Le Bris. "Domain Decomposition for Electronic Structure Computations". In: *Parallel Computing: Architectures, Algorithms and Applications*. IOS Press, 2008, pp. 29–36 (cit. on p. 76).

[35] P. Bastian, M. Blatt, A. Dedner, N.-A. Dreier, C. Engwer, R. Fritze, C. Gräser, C. Grüninger, D. Kempf, R. Klöfkorn, M. Ohlberger, and O. Sander. "The Dune Framework: Basic Concepts and Recent Developments". In: *Computers & Mathematics with Applications*. Development and Application of Open-source Software for Problems with Numerical PDEs 81 (2021), pp. 75–112. DOI: 10.1016/j.camwa.2020.06.007 (cit. on p. 22).

[36] P. Bastian, R. Scheichl, L. Seelinger, and A. Strehlow. "Multilevel Spectral Domain Decomposition". In: *SIAM J. Sci. Comput.* (2022), S1–S26. DOI: 10.1137/21M1427231 (cit. on pp. 37, 77, 84, 87, 89–92).

[37] P. F. Baumeister. "Real-Space Finite-Difference PAW Method for Large-Scale Applications on Massively Parallel Computers". PhD Thesis. Forschungszentrum Jülich, 2013 (cit. on p. 20).

[38] T. L. Beck. "Real-Space Mesh Techniques in Density-Functional Theory". In: *Rev. Mod. Phys.* 72.4 (2000), pp. 1041–1080. DOI: 10.1103/RevModPhys.72.1041 (cit. on pp. 19, 20).

[39] G. Bencteux, E. Cancés, W. W. Hager, and C. L. Bris. "Analysis of a Quadratic Programming Decomposition Algorithm". In: *SIAM J. Numer. Anal.* 47.6 (2010), pp. 4517–4539. DOI: 10.1137/070701728 (cit. on p. 77).

[40] G. Bencteux, M. Barrault, E. Cancès, W. W. Hager, and C. Le Bris. "Domain Decomposition and Electronic Structure Computations: A Promising Approach". In: *Partial Differential Equations: Modeling and Numerical Simulation*. Ed. by R. Glowinski and P. Neittaanmäki. Computational Methods in Applied Sciences. Dordrecht: Springer Netherlands, 2008, pp. 147–164. DOI: 10.1007/978-1-4020-8758-5_8 (cit. on p. 77).

[41] P. Benner and T. Mach. "Locally Optimal Block Preconditioned Conjugate Gradient Method for Hierarchical Matrices". In: *PAMM* 11.1 (2011), pp. 741–742. DOI: 10.1002/pamm.201110360 (cit. on p. 93).

[42] J. K. Bennighof and R. B. Lehoucq. "An Automated Multilevel Substructuring Method for Eigenspace Computation in Linear Elastodynamics". In: *SIAM J. Sci. Comput.* 25.6 (2004), pp. 2084–2106. DOI: 10.1137/S1064827502400650 (cit. on p. 77).

[43] S. Berger. "Density Operator in Eigenvalue Problems with Application in Manifold Interpolation". Bachelor Thesis. RWTH Aachen University, 2022 (cit. on p. xi).

[44] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah. "Julia: A Fresh Approach to Numerical Computing". In: *SIAM Rev.* 59.1 (2017), pp. 65–98. DOI: 10.1137/141000671 (cit. on pp. 60, 93).

[45] N. Bootland, V. Dolean, I. G. Graham, C. Ma, and R. Scheichl. "Overlapping Schwarz Methods with GenEO Coarse Spaces for Indefinite and Nonself-Adjoint Problems". In: *IMA J. Numer. Anal.* 43.4 (2023), pp. 1899–1936. DOI: 10.1093/imanum/drac036 (cit. on p. 87).

[46] H. Borchardt. "Iterative Domain Decomposition Methods for Eigenvalue Problems". Master Thesis. RWTH Aachen University, 2020 (cit. on pp. xii, 2).

[47] M. Born and R. Oppenheimer. "Zur Quantentheorie der Molekeln". In: *Annalen der Physik* 389.20 (1927), pp. 457–484. DOI: 10.1002/andp.19273892002 (cit. on p. 10).

[48] J. H. Bramble, J. E. Pasciak, and A. V. Knyazev. "A Subspace Preconditioning Algorithm for Eigenvector/Eigenvalue Computation". In: *Adv Comput Math* 6.1 (1996), pp. 159–189. DOI: 10.1007/BF02127702 (cit. on p. 30).

[49] A. Brand, L. Allen, M. Altman, M. Hlava, and J. Scott. "Beyond Authorship: Attribution, Contribution, Collaboration, and Credit". In: *Learned Publishing* 28.2 (2015), pp. 151–155. DOI: 10.1087/20150211 (cit. on p. xiii).

[50] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. Vol. 15. Texts in Applied Mathematics. New York: Springer, 1994. DOI: 10.1007/978-1-4757-4338-8 (cit. on p. 77).

[51] S. C. Brenner and L. R. Scott. *The Mathematical Theory of Finite Element Methods*. 3. ed. Texts in Applied Mathematics 15. New York: Springer, 2008. DOI: 10.1007/978-0-387-75934-0 (cit. on p. 21).

[52] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. New York: Springer, 2010. DOI: 10.1007/978-0-387-70914-7 (cit. on p. 49).

[53] F. Brezzi and M. Fortin. *Mixed and Hybrid Finite Element Methods*. Springer Series in Computational Mathematics 15. New York, Berlin, Heidelberg: Springer, 1991. DOI: 10.1007/978-1-4612-3172-1 (cit. on pp. 16, 21).

[54] X.-C. Cai and M. Sarkis. "A Restricted Additive Schwarz Preconditioner for General Sparse Linear Systems". In: *SIAM J. Sci. Comput.* 21.2 (1999), pp. 792–797. DOI: 10.1137/S106482759732678X (cit. on p. 80).

[55] M. Caliari, A. Ostermann, S. Rainer, and M. Thalhammer. "A Minimisation Approach for Computing the Ground State of Gross–Pitaevskii Systems". In: *Journal of Computational Physics* 228.2 (2009), pp. 349–360. DOI: 10.1016/j.jcp.2008.09.018 (cit. on p. 23).

[56] E. Cancès. "Introduction to First-Principle Simulation of Molecular Systems". In: *Computational Mathematics, Numerical Analysis and Applications: Lecture Notes of the XVII 'Jacques-Louis Lions' Spanish-French School*. Ed. by M. Mateos and P. Alonso. SEMA SIMAI Springer Series. Cham: Springer International Publishing, 2017, pp. 61–106. DOI: 10.1007/978-3-319-49631-3_2 (cit. on pp. 7–12, 14).

[57] E. Cancès. "Self-Consistent Field (SCF) Algorithms". In: *Encyclopedia of Applied and Computational Mathematics*. Ed. by B. Engquist. Berlin, Heidelberg: Springer, 2015, pp. 1310–1316. DOI: 10.1007/978-3-540-70529-1_256 (cit. on p. 22).

[58] E. Cancès and C. L. Bris. "On the Convergence of SCF Algorithms for the Hartree-Fock Equations". In: *ESAIM: M2AN* 34.4 (2000), pp. 749–774. DOI: 10.1051/m2an:2000102 (cit. on p. 23).

[59] E. Cancès, M. Defranceschi, W. Kutzelnigg, C. Le Bris, and Y. Maday. "Computational Quantum Chemistry: A Primer". In: *Handbook of Numerical Analysis*. Vol. 10. Special Volume, Computational Chemistry. Elsevier, 2003, pp. 3–270. DOI: 10.1016/S1570-8659(03)10003-8 (cit. on pp. 10–18, 20, 23).

[60] E. Cancès and G. Friesecke, eds. *Density Functional Theory: Modeling, Mathematical Analysis, Computational Methods, and Applications*. Cham: Springer, 2023. DOI: 10.1007/978-3-031-22340-2 (cit. on pp. 16, 22, 76).

[61] E. Cancès, L. Garrigue, and D. Gontier. *Second-Order Homogenization of Periodic Schrödinger Operators with Highly Oscillating Potentials*. 2021. arXiv: 2112.12008 (cit. on p. 42).

[62] E. Cancès, L. Garrigue, and D. Gontier. "Simple Derivation of Moiré-Scale Continuous Models for Twisted Bilayer Graphene". In: *Phys. Rev. B* 107.15 (2023), 155403:1–155403:12. DOI: 10.1103/PhysRevB.107.155403 (cit. on p. 76).

[63] E. Cancès, M. Hassan, and L. Vidal. *Modified-Operator Method for the Calculation of Band Diagrams of Crystalline Materials*. 2022. arXiv: 2210.00442 [cond-mat] (cit. on p. 76).

[64] E. Cancès, G. Kemlin, and A. Levitt. "Convergence Analysis of Direct Minimization and Self-Consistent Iterations". In: *SIAM J. Matrix Anal. Appl.* 42.1 (2021), pp. 243–274. DOI: 10.1137/20M1332864 (cit. on pp. 42, 76).

[65] E. Cancès, Y. Maday, and B. Stamm. "Domain Decomposition for Implicit Solvation Models". In: *J. Chem. Phys.* 139.5 (2013). DOI: 10.1063/1.4816767 (cit. on pp. 2, 76, 96).

[66] Y. Cao, V. Fatemi, A. Demir, S. Fang, S. L. Tomarken, J. Y. Luo, J. D. Sanchez-Yamagishi, K. Watanabe, T. Taniguchi, E. Kaxiras, R. C. Ashoori, and P. Jarillo-Herrero. "Correlated Insulator Behaviour at Half-Filling in Magic-Angle Graphene Superlattices". In: *Nature* 556.7699 (2018), pp. 80–84. DOI: 10.1038/nature26154 (cit. on p. 76).

[67] Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero. "Unconventional Superconductivity in Magic-Angle Graphene Superlattices". In: *Nature* 556.7699 (2018), pp. 43–50. DOI: 10.1038/nature26160 (cit. on p. 76).

[68] S. Capozziello and W.-G. Boskoff. *A Mathematical Journey to Quantum Mechanics*. Unitext for Physics. Cham: Springer International Publishing, 2021. DOI: 10.1007/978-3-030-86098-1 (cit. on p. 8).

[69] S. Carr, D. Massatt, S. B. Torrisi, P. Cazeaux, M. Luskin, and E. Kaxiras. "Relaxation and Domain Formation in Incommensurate Two-Dimensional Heterostructures". In: *Phys. Rev. B* 98.22 (2018), pp. 224102-1–224102-7. DOI: 10.1103/PhysRevB.98.224102 (cit. on p. 42).

[70] P. Cazeaux, M. Luskin, and D. Massatt. "Energy Minimization of Two Dimensional Incommensurate Heterostructures". In: *Arch Rational Mech Anal* 235.2 (2020), pp. 1289–1325. DOI: 10.1007/s00205-019-01444-y (cit. on p. 42).

[71] F. Chatelin. *Eigenvalues of Matrices*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2012. DOI: 10.1137/1.9781611972467 (cit. on p. 26).

[72] H. Chen, X. Dai, X. Gong, L. He, and A. Zhou. "Adaptive Finite Element Approximations for Kohn–Sham Models". In: *Multiscale Model. Simul.* 12.4 (2014), pp. 1828–1869. DOI: 10.1137/130916096 (cit. on p. 22).

[73] H. Chen and C. Ortner. "QM/MM Methods for Crystalline Defects. Part 1: Locality of the Tight Binding Model". In: *Multiscale Model. Simul.* 14.1 (2016), pp. 232–264. DOI: 10.1137/15M1022628 (cit. on p. 103).

[74] J. Chen and J. Lu. "Analysis of the Divide-and-Conquer Method for Electronic Structure Calculations". In: *Math. Comp.* 85.302 (2016), pp. 2919–2938. DOI: 10.1090/mcom/3066 (cit. on p. 77).

[75] C. Chevalier and F. Pellegrini. "PT-Scotch: A Tool for Efficient Parallel Graph Ordering". In: *Parallel Comput.* Parallel Matrix Algorithms and Applications 34.6 (2008), pp. 318–331. DOI: 10.1016/j.parco.2007.12.001 (cit. on pp. 34, 83).

[76] M. Chipot and A. Rougirel. "On the Asymptotic Behaviour of the Solution of Elliptic Problems in Cylindrical Domains Becoming Unbounded". In: *Commun. Contemp. Math.* 04.01 (2002), pp. 15–44. DOI: 10.1142/S0219199702000555 (cit. on p. 42).

[77] M. Chipot. "L goes to plus Infinity: An Update". In: *Journal of the Korean Society for Industrial and Applied Mathematics* 18.2 (2014), pp. 107–127. DOI: 10.12941/JKSIAM.2014.18.107 (cit. on p. 42).

[78] M. Chipot. "On Some Anisotropic Singular Perturbation Problems". In: *Asymptotic Analysis* (2007), p. 21. DOI: 10.5167/uzh-21524 (cit. on p. 52).

[79] M. Chipot, A. Elfanni, and A. Rougirel. "Eigenvalues, Eigenfunctions in Domains Becoming Unbounded". In: *Hyperbolic Problems and Regularity Questions*. Ed. by M. Padula and L. Zanghirati. Basel: Birkhäuser, 2007, pp. 69–78. DOI: 10.1007/978-3-7643-7451-8_8 (cit. on p. 42).

[80] M. Chipot, W. Hackbusch, S. Sauter, and A. Veit. "Numerical Approximation of Poisson Problems in Long Domains". In: *Vietnam J. Math.* (2021). DOI: 10.1007/s10013-021-00512-9 (cit. on p. 42).

[81] M. Chipot and A. Rougirel. "On the Asymptotic Behaviour of the Eigenmodes for Elliptic Problems in Domains Becoming Unbounded". In: *Trans. Amer. Math. Soc.* 360.07 (2008), pp. 3579–3603. DOI: 10.1090/S0002-9947-08-04361-4 (cit. on p. 42).

[82] M. Chipot, P. Roy, and I. Shafrir. "Asymptotics of Eigenstates of Elliptic Problems with Mixed Boundary Data on Domains Tending to Infinity". In: *Asymptotic Analysis* 85.3-4 (2013), pp. 199–227. DOI: 10.3233/ASY-131182 (cit. on p. 42).

[83] M. Chipot and Y. Xie. "On the Asymptotic Behaviour of Elliptic Problems with Periodic Data". In: *Comptes Rendus Mathematique* 339.7 (2004), pp. 477–482. DOI: 10.1016/j.crma.2004.09.007 (cit. on p. 42).

[84] R. E. Christoffersen. *Basic Principles and Techniques of Molecular Quantum Mechanics*. Ed. by C. R. Cantor. Springer Advanced Texts in Chemistry. New York, NY: Springer US, 1989. DOI: 10.1007/978-1-4684-6360-6 (cit. on pp. 7, 9).

[85] M. Chupeng, C. Alber, and R. Scheichl. "Wavenumber Explicit Convergence of a Multiscale Generalized Finite Element Method for Heterogeneous Helmholtz Problems". In: *SIAM J. Numer. Anal.* 61.3 (2023), pp. 1546–1584. DOI: 10.1137/21M1466748 (cit. on p. 77).

[86] G. Ciaramella and M. J. Gander. "Analysis of the Parallel Schwarz Method for Growing Chains of Fixed-Sized Subdomains: Part I". In: *SIAM J. Numer. Anal.* 55.3 (2017), pp. 1330–1356. DOI: 10.1137/16M1065215 (cit. on pp. 2, 76).

[87] G. Ciaramella and M. J. Gander. "Analysis of the Parallel Schwarz Method for Growing Chains of Fixed-sized Subdomains: Part II". In: *SIAM J. Numer. Anal.* 56.3 (2018), pp. 1498–1524. DOI: 10.1137/17M1115885 (cit. on pp. 2, 76).

[88] G. Ciaramella and M. J. Gander. "Analysis of the Parallel Schwarz Method for Growing Chains of Fixed-Sized Subdomains: Part III". In: *ETNA* 49 (2018), pp. 210–243. DOI: 10.1553/etna_vol49s210 (cit. on pp. 2, 76).

[89] G. Ciaramella and M. J. Gander. *Iterative Methods and Preconditioners for Systems of Linear Equations*. SIAM, 2022. DOI: 10.1007/978-3-0348-7893-7 (cit. on pp. 32, 35, 36, 73, 80).

[90] G. Ciaramella, M. Hassan, and B. Stamm. "On the Scalability of the Parallel Schwarz Method in One-Dimension". In: *Domain Decomposition Methods in Science and Engineering XXV*. Ed. by R. Haynes, S. MacLachlan, X.-C. Cai, L. Halpern, H. H. Kim, A. Klawonn, and O. Widlund. Lecture Notes in Computational Science and Engineering. Cham: Springer International Publishing, 2020, pp. 151–158. DOI: 10.1007/978-3-030-56750-7_16 (cit. on pp. 2, 76).

[91] G. Ciaramella, M. Hassan, and B. Stamm. "On the Scalability of the Schwarz Method". In: *SMAI J. Comput. Math.* 6 (2020), pp. 33–68. DOI: 10.5802/smai-jcm.61 (cit. on pp. 2, 76).

[92] C. Cohen-Tannoudji, B. Diu, and F. Laloë. *Quantum Mechanics*. Trans. by S. R. Hemley, N. Ostrowsky, and D. B. Ostrowsky. Second edition. Vol. 1. Weinheim: Wiley-VCH Verlag GmbH & Co, 2020 (cit. on p. 8).

[93] R. Courant and D. Hilbert. *Methods of Mathematical Physics. Volume I*. New York: Wiley, 1989. DOI: 10.1002/9783527617210 (cit. on p. 45).

[94] A. St-Cyr, M. J. Gander, and S. J. Thomas. "Optimized Multiplicative, Additive, and Restricted Additive Schwarz Preconditioning". In: *SIAM J. Sci. Comput.* 29.6 (2007), pp. 2402–2425. DOI: 10.1137/060652610 (cit. on p. 34).

[95] M. Defranceschi and C. Le Bris. *Mathematical Models and Methods for Ab Initio Quantum Chemistry*. Ed. by G. Berthier, H. Fischer, K. Fukui, G. G. Hall, J. Hinze, J. Jortner, W. Kutzelnigg, K. Ruedenberg, and J. Tomasi. Vol. 74. Lecture Notes in Chemistry. Berlin, Heidelberg: Springer, 2000. DOI: 10.1007/978-3-642-57237-1 (cit. on pp. 7, 9, 14, 23).

[96] J. W. Demmel. *Applied Numerical Linear Algebra*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 1997. DOI: 10.1137/1.9781611971446 (cit. on p. 26).

[97] V. Dolean, M. J. Gander, W. Kheriji, F. Kwok, and R. Masson. "Nonlinear Preconditioning: How to Use a Nonlinear Schwarz Method to Precondition Newton's Method". In: *SIAM J. Sci. Comput.* 38.6 (2016), A3357–A3380. DOI: 10.1137/15M102887X (cit. on p. 77).

[98] V. Dolean, P. Jolivet, and F. Nataf. *An Introduction to Domain Decomposition Methods*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2015. DOI: 10.1137/1.9781611974065 (cit. on pp. 32–36, 73, 80–84, 90, 93, 95).

[99] V. Dolean, F. Nataf, R. Scheichl, and N. Spillane. "Analysis of a Two-level Schwarz Method with Coarse Spaces Based on Local Dirichlet-to-Neumann Maps". In: *Comput. Methods Appl. Math.* 12.4 (2012), pp. 391–414. DOI: 10.2478/cmam-2012-0027 (cit. on pp. 77, 85).

[100] R. Dong, D. Li, and L. Wang. "Directional Homogenization of Elliptic Equations in Non-Divergence Form". In: *Journal of Differential Equations* 268.11 (2020), pp. 6611–6645. DOI: 10.1016/j.jde.2019.11.041 (cit. on p. 42).

[101] R. Dong, D. Li, and L. Wang. "Regularity of Elliptic Systems in Divergence Form with Directional Homogenization". In: *Discrete & Continuous Dynamical Systems* 38.1 (2018), pp. 75–90. DOI: 10.3934/dcds.2018004 (cit. on p. 42).

[102] B. Dörich and P. Henning. *Error Bounds for Discrete Minimizers of the Ginzburg-Landau Energy in the High-κ Regime*. 2023. arXiv: 2303.13961 [cs, math] (cit. on p. 76).

[103] M. Dryja and O. Widlund. *An Additive Variant of the Schwarz Alternating Method for the Case of Many Subregions*. Tech. rep. Courant Institute, 1987 (cit. on p. 80).

[104] M.-S. Dupuy. "Analyse de la méthode projector augmented-wave pour les calculs de structure électronique en géométrie périodique". PhD Thesis. Université Sorbonne Paris Cité, 2018 (cit. on pp. 13, 15).

[105] G. Dusson. "Estimation d'erreur Pour Des Problèmes Aux Valeurs Propres Linéaires et Non-Linéaires Issus Du Calcul de Structure Électronique". Thèse de Doctorat. Université Pierre et Marie Curie - Paris VI, 2017 (cit. on pp. 10–12, 15, 20).

[106] A. Edelman, T. A. Arias, and S. T. Smith. "The Geometry of Algorithms with Orthogonality Constraints". In: *SIAM J. Matrix Anal. & Appl.* 20.2 (1998), pp. 303–353. DOI: 10.1137/S0895479895290954 (cit. on p. 23).

[107] J.-L. Fattebert and J. Bernholc. "Towards Grid-Based O(N) Density-Functional Theory Methods: Optimized Nonorthogonal Orbitals and Multigrid Acceleration". In: *Phys. Rev. B* 62.3 (2000), pp. 1713–1722. DOI: 10.1103/PhysRevB.62.1713 (cit. on p. 20).

[108] J.-L. Fattebert. "Finite Difference Schemes and Block Rayleigh Quotient Iteration for Electronic Structure Calculations on Composite Grids". In: *Journal of Computational Physics* 149.1 (1999), pp. 75–94. DOI: 10.1006/jcph.1998.6138 (cit. on p. 20).

[109] J.-L. Fattebert and M. B. Nardelli. "Finite Difference Methods for Ab Initio Electronic Structure and Quantum Transport Calculations of Nanostructures". In: *Handbook of Numerical Analysis*. Vol. 10. Special Volume, Computational Chemistry. Elsevier, 2003, pp. 571–612. DOI: 10.1016/S1570-8659(03)10009-9 (cit. on p. 20).

[110] Y. T. Feng and D. R. J. Owen. "Conjugate Gradient Methods For Solving The Smallest Eigenpair Of Large Symmetric Eigenvalue Problems". In: *Int. J. Numer. Meth. Engng.* 39.13 (1996), pp. 2209–2229. DOI: 10.1002/(SICI)1097-0207(19960715)39:13<2209::AID-NME951>3.0.CO;2-R (cit. on p. 32).

[111] E. Fermi. "Un Metodo Statistico per La Determinazione Di Alcune Priorieta Dell'atome". In: *Rend. Accad. Naz. Lincei* 6 (1927), pp. 602–607 (cit. on p. 16).

[112] L. Fiorenza. "Preconditioning in Steepest Descent Methods for Discretized Elliptic Eigenvalue Problems". Ongoing. Bachelor Thesis. University of Stuttgart, 2024 (cit. on p. xi).

[113] J. B. Francisco, J. M. Martínez, and L. Martínez. "Globally Convergent Trust-Region Methods for Self-Consistent Field Electronic Structure Calculations". In: *The Journal of Chemical Physics* 121.22 (2004), pp. 10863–10878. DOI: 10.1063/1.1814935 (cit. on p. 23).

[114] M. A. Freitag and A. Spence. "Convergence of Inexact Inverse Iteration with Application to Preconditioned Iterative Solves". In: *BIT* 47.1 (2007), pp. 27–44. DOI: 10.1007/s10543-006-0100-1 (cit. on pp. 28, 42, 79, 97).

[115] M. A. Freitag. "Inner-Outer Iterative Methods for Eigenvalue Problems - Convergence and Preconditioning". PhD Thesis. University of Bath, 2007 (cit. on pp. 28, 79).

[116] N. Friess, A. D. Gilbert, and R. Scheichl. *A Complex-Projected Rayleigh Quotient Iteration for Targeting Interior Eigenvalues*. 2023. DOI: 10.48550/arXiv.2312.02847. arXiv: 2312.02847 [cs, math] (cit. on p. 29).

[117] J. Gaidamour, Q. Tang, and X. Antoine. "BEC2HPC: A HPC Spectral Solver for Nonlinear Schrödinger and Rotating Gross-Pitaevskii Equations. Stationary States Computation". In: *Comput. Phys. Commun.* 265 (2021), p. 108007. DOI: 10.1016/j.cpc.2021.108007 (cit. on p. 76).

[118] M. Galewski, B. Galewska, and E. Schmeidel. "Conditions for Having a Diffeomorphism between Two Banach Spaces". In: *Electronic Journal of Differential Equations* 99 (2014), pp. 1–6 (cit. on p. 49).

[119] J. Galvis and Y. Efendiev. "Domain Decomposition Preconditioners for Multiscale Flows in High Contrast Media: Reduced Dimension Coarse Spaces". In: *Multiscale Model. Simul.* 8.5 (2010), pp. 1621–1644. DOI: 10.1137/100790112 (cit. on p. 77).

[120] J. Galvis and Y. Efendiev. "Domain Decomposition Preconditioners for Multiscale Flows in High-Contrast Media". In: *Multiscale Model. Simul.* 8.4 (2010), pp. 1461–1483. DOI: 10.1137/090751190 (cit. on p. 77).

[121] M. J. Gander. "Optimized Schwarz Methods". In: *SIAM J. Numer. Anal.* 44.2 (2006), pp. 699–731. DOI: 10.1137/S0036142903425409 (cit. on p. 34).

[122] M. J. Gander. "Schwarz Methods over the Course of Time." In: *ETNA. Electronic Transactions on Numerical Analysis* 31 (2008), pp. 228–255 (cit. on pp. 32, 33).

[123] M. J. Gander and A. Loneland. "SHEM: An Optimal Coarse Space for RAS and Its Multiscale Approximation". In: *Domain Decomposition Methods in Science and Engineering XXIII*. Ed. by C.-O. Lee, X.-C. Cai, D. E. Keyes, H. H. Kim, A. Klawonn, E.-J. Park, and O. B. Widlund. Lecture Notes in Computational Science and Engineering. Cham: Springer International Publishing, 2017, pp. 313–321. DOI: 10.1007/978-3-319-52389-7_32 (cit. on p. 37).

[124] M. J. Gander and B. Song. "Complete, Optimal and Optimized Coarse Spaces for Additive Schwarz". In: *Domain Decomposition Methods in Science and Engineering XXIV*. Ed. by P. E. Bjørstad, S. C. Brenner, L. Halpern, H. H. Kim, R. Kornhuber, T. Rahman, and O. B. Widlund. Lecture Notes in Computational Science and Engineering. Cham: Springer International Publishing, 2018, pp. 301–309. DOI: 10.1007/978-3-319-93873-8_28 (cit. on p. 80).

[125] L. Genovese, A. Neelov, S. Goedecker, T. Deutsch, S. A. Ghasemi, A. Willand, D. Caliste, O. Zilberberg, M. Rayson, A. Bergman, and R. Schneider. "Daubechies Wavelets as a Basis Set for Density Functional Pseudopotential Calculations". In: *The Journal of Chemical Physics* 129.1 (2008). DOI: 10.1063/1.2949547 (cit. on p. 20).

[126] E. Gerstner. "Nobel Prize 2010: Andre Geim & Konstantin Novoselov". In: *Nat. Phys.* 6.11 (2010), pp. 836–836. DOI: 10.1038/nphys1836 (cit. on pp. 24, 76).

[127] D. Gilbarg and N. S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. 2nd ed. Classics in Mathematics. Berlin Heidelberg: Springer, 2001. DOI: 10.1007/978-3-642-61798-0 (cit. on pp. 44, 88).

[128] G. H. Golub and H. A. van der Vorst. "Eigenvalue Computation in the 20th Century". In: *J. Comput. Appl. Math.* Numerical Analysis 2000. Vol. III: Linear Algebra 123.1 (2000), pp. 35–65. DOI: 10.1016/S0377-0427(00)00413-1 (cit. on p. 26).

[129] G. H. Golub and C. F. Van Loan. *Matrix Computations*. Fourth edition. Baltimore: The Johns Hopkins University Press, 2013. DOI: 10.56021/9781421407944 (cit. on pp. 26, 74).

[130] L. Gouarin and N. Spillane. *Fully Algebraic Domain Decomposition Preconditioners with Adaptive Spectral Bounds*. 2021. arXiv: 2106.10913 [cs, math] (cit. on pp. 37, 77).

[131] D. S. Grebenkov and B.-T. Nguyen. "Geometrical Structure of Laplacian Eigenfunctions". In: *SIAM Rev.* 55.4 (2013), pp. 601–667. DOI: 10.1137/120880173 (cit. on p. 103).

[132] D. J. Griffiths and D. F. Schroeter. *Introduction to Quantum Mechanics*. 3rd ed. Cambridge University Press, 2018. DOI: 10.1017/9781316995433 (cit. on p. 8).

[133] S. J. Gustafson and I. M. Sigal. *Mathematical Concepts of Quantum Mechanics*. Universitext. Berlin, Heidelberg: Springer, 2011. DOI: 10.1007/978-3-642-21866-8 (cit. on pp. 7, 18, 19).

[134] W. Hackbusch. *Theorie und Numerik elliptischer Differentialgleichungen*. Wiesbaden: Springer, 2017. DOI: 10.1007/978-3-658-15358-8 (cit. on p. 19).

[135] D. R. Hartree. *The Calculation of Atomic Structures*. J. Wiley, 1957 (cit. on p. 13).

[136] M. Hassan. "Mathematical Analysis of Boundary Integral Equations and Domain Decomposition Methods with Applications in Polarisable Electrostatics". PhD Thesis. RWTH Aachen University, 2020. DOI: 10.18154/RWTH-2020-07206 (cit. on pp. 2, 76).

[137] 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: *Numer. Math.* 155.1 (2023), pp. 121–173. DOI: 10.1007/s00211-023-01371-x (cit. on p. 14).

[138] F. Hecht. "New Development in Freefem++". In: *Journal of Numerical Mathematics* 20.3-4 (2012), pp. 251–266. DOI: 10.1515/jnum-2012-0013 (cit. on p. 22).

[139] P. Heid, B. Stamm, and T. P. Wihler. "Gradient Flow Finite Element Discretizations with Energy-Based Adaptivity for the Gross-Pitaevskii Equation". In: *J. Comput. Phys.* 436 (2021), p. 110165. DOI: 10.1016/j.jcp.2021.110165 (cit. on pp. 23, 42, 76).

[140] A. Heinlein, A. Klawonn, J. Knepper, and O. Rheinbach. "Adaptive GDSW Coarse Spaces for Overlapping Schwarz Methods in Three Dimensions". In: *SIAM J. Sci. Comput.* 41.5 (2019), A3045–A3072. DOI: 10.1137/18M1220613 (cit. on pp. 37, 77).

[141] A. Heinlein, A. Klawonn, J. Knepper, and O. Rheinbach. "Multiscale coarse spaces for overlapping Schwarz methods based on the ACMS space in 2D". In: *ETNA* 48 (2018), pp. 156–182. DOI: 10.1553/etna_vol48s156 (cit. on pp. 37, 77).

[142] T. Helgaker, P. Jørgensen, and J. Olsen. *Molecular Electronic-Structure Theory*. Chichester ; New York: Wiley, 2000 (cit. on p. 12).

[143] P. Henning. "The Dependency of Spectral Gaps on the Convergence of the Inverse Iteration for a Nonlinear Eigenvector Problem". In: *Math. Models Methods Appl. Sci.* 33.07 (2023), pp. 1517–1544. DOI: 10.1142/S0218202523500343 (cit. on p. 76).

[144] P. Henning, A. Målqvist, and D. Peterseim. "Two-Level Discretization Techniques for Ground State Computations of Bose-Einstein Condensates". In: *SIAM J. Numer. Anal.* 52.4 (2014), pp. 1525–1550. DOI: 10.1137/130921520 (cit. on p. 77).

[145] P. Henning and A. Persson. "On Optimal Convergence Rates for Discrete Minimizers of the Gross–Pitaevskii Energy in Localized Orthogonal Decomposition Spaces". In: *Multiscale Model. Simul.* (2023), pp. 993–1011. DOI: 10.1137/22M1516300 (cit. on p. 77).

[146] P. Henning and D. Peterseim. "Sobolev Gradient Flow for the Gross–Pitaevskii Eigenvalue Problem: Global Convergence and Computational Efficiency". In: *SIAM J. Numer. Anal.* 58.3 (2020), pp. 1744–1772. DOI: 10.1137/18M1230463 (cit. on pp. 23, 42, 44, 59, 66, 76).

[147] A. Henrot. *Extremum Problems for Eigenvalues of Elliptic Operators*. Frontiers in Mathematics. Birkhäuser Basel, 2006. DOI: 10.1007/3-7643-7706-2 (cit. on p. 67).

[148] M. F. Herbst and A. Levitt. "Black-Box Inhomogeneous Preconditioning for Self-Consistent Field Iterations in Density Functional Theory". In: *J. Phys.: Condens. Matter* 33.8 (2021), p. 085503. DOI: 10.1088/1361-648X/abcbdb (cit. on p. 42).

[149] M. F. Herbst and A. Levitt. "A Robust and Efficient Line Search for Self-Consistent Field Iterations". In: *J. Comput. Phys.* 459 (2022), p. 111127. DOI: 10.1016/j.jcp.2022.111127 (cit. on p. 76).

[150] M. F. Herbst, A. Levitt, and E. Cancès. "DFTK: A Julian Approach for Simulating Electrons in Solids". In: *Proceedings of the JuliaCon Conferences* 3.26 (2021), p. 69. DOI: 10.21105/jcon.00069 (cit. on p. 76).

[151] M. Hestenes and E. Stiefel. "Methods of Conjugate Gradients for Solving Linear Systems". In: *J. Res. Natl. Inst. Stand. Technol.* 49.6 (1952), pp. 409–436. DOI: 10.6028/jres.049.044 (cit. on p. 73).

[152] K.-H. Hoffmann and J. Zou. "Parallel Efficiency of Domain Decomposition Methods". In: *Parallel Computing* 19.12 (1993), pp. 1375–1391. DOI: 10.1016/0167-8191(93)90082-V (cit. on p. 2).

[153] P. Hohenberg and W. Kohn. "Inhomogeneous Electron Gas". In: *Phys. Rev.* 136.3B (1964), B864–B871. DOI: 10.1103/PhysRev.136.B864 (cit. on p. 16).

[154] S. Høst, J. Olsen, B. Jansík, L. Thøgersen, P. Jørgensen, and T. Helgaker. "The Augmented Roothaan–Hall Method for Optimizing Hartree–Fock and Kohn–Sham Density Matrices". In: *The Journal of Chemical Physics* 129.12 (2008), p. 124106. DOI: 10.1063/1.2974099 (cit. on p. 23).

[155] A. S. Householder. *The Theory of Matrices in Numerical Analysis*. Blaisdell, 1964 (cit. on p. 26).

[156] W. Hu, L. Lin, and C. Yang. "DGDFT: A Massively Parallel Method for Large Scale Density Functional Theory Calculations". In: *The Journal of Chemical Physics* 143.12 (2015), p. 124110. DOI: 10.1063/1.4931732 (cit. on p. 22).

[157] S. Iijima. "Helical Microtubules of Graphitic Carbon". In: *Nature* 354.6348 (1991), pp. 56–58. DOI: 10.1038/354056a0 (cit. on p. 76).

[158] V. V. Jikov, S. M. Kozlov, and O. A. Oleinik. *Homogenization of Differential Operators and Integral Functionals*. Berlin, Heidelberg: Springer-Verlag, 1994. DOI: 10.1007/978-3-642-84659-5 (cit. on p. 58).

[159] V. Kalantzis, J. Kestyn, E. Polizzi, and Y. Saad. "Domain Decomposition Approaches for Accelerating Contour Integration Eigenvalue Solvers for Symmetric Eigenvalue Problems: Domain Decomposition Contour Integration Eigensolvers". In: *Numer. Linear Algebra Appl.* 25.5 (2018), pp. 1–19. DOI: 10.1002/nla.2154 (cit. on p. 77).

[160] H. Karamitaheri. "Thermal and Thermoelectric Properties of Nanostructures". PhD Thesis. Technische Universität Wien, 2013. DOI: 10.34726/hss.2013.29976 (cit. on p. 103).

[161] G. Karypis and V. Kumar. *METIS: A Software Package for Partitioning Unstructured Graphs, Partitioning Meshes, and Computing Fill-Reducing Orderings of Sparse Matrices*. Tech. rep. University of Minnesota, 1997, pp. 1–31 (cit. on pp. 34, 83).

[162] C. T. Kelley. *Iterative Methods for Linear and Nonlinear Equations*. SIAM, 1995. DOI: 10.1137/1.9781611970944 (cit. on p. 74).

[163] G. Kemlin. "Analyse Numérique Pour La Théorie de La Fonctionnelle de Densité". Thèse de Doctorat. École des ponts ParisTech, 2022 (cit. on p. 13).

[164] S. Kesavan. "Homogenization of Elliptic Eigenvalue Problems: Part 1". In: *Appl Math Optim* 5.1 (1979), pp. 153–167. DOI: 10.1007/BF01442551 (cit. on pp. 42, 45, 52, 56).

[165] S. Kesavan. "Homogenization of Elliptic Eigenvalue Problems: Part 2". In: *Appl Math Optim* 5.1 (1979), pp. 197–216. DOI: 10.1007/BF01442554 (cit. on pp. 42, 52, 56).

[166] A. Klawonn, M. Lanser, and O. Rheinbach. "Toward Extremely Scalable Nonlinear Domain Decomposition Methods for Elliptic Partial Differential Equations". In: *SIAM J. Sci. Comput.* 37.6 (2015), pp. C667–C696. DOI: 10.1137/140997907 (cit. on p. 79).

[167] A. V. Knyazev. "A Preconditioned Conjugate Gradient Method for Eigenvalue Problems and Its Implementation in a Subspace". In: *Numerical Treatment of Eigenvalue Problems Vol. 5*. Ed. by K.-H. Hoffmann, H. D. Mittelmann, J. Todd, J. Albrecht, L. Collatz, P. Hagedorn, and W. Velte. Vol. 96. Basel: Birkhäuser, 1991, pp. 143–154. DOI: 10.1007/978-3-0348-6332-2_11 (cit. on pp. 32, 72, 79).

[168] A. V. Knyazev. “Convergence Rate Estimates for Iterative Methods for a Mesh Symmetrie Eigenvalue Problem”. In: *Russian Journal of Numerical Analysis and Mathematical Modelling* 2.5 (1987), pp. 371–396. DOI: 10.1515/rnam.1987.2.5.371 (cit. on p. 30).

[169] A. V. Knyazev and A. L. Shorokhodov. “On Exact Estimates of the Convergence Rate of the Steepest Ascent Method in the Symmetric Eigenvalue Problem”. In: *Linear Algebra and its Applications* 154–156 (1991), pp. 245–257. DOI: 10.1016/0024-3795(91)90379-B (cit. on p. 30).

[170] A. V. Knyazev. “Preconditioned Eigensolvers—An Oxymoron?” In: *ETNA* 7 (1998), pp. 104–123 (cit. on pp. 28, 30).

[171] A. V. Knyazev. “Toward the Optimal Preconditioned Eigensolver: Locally Optimal Block Preconditioned Conjugate Gradient Method”. In: *SIAM J. Sci. Comput.* 23.2 (2001), pp. 517–541. DOI: 10.1137/S1064827500366124 (cit. on p. 32).

[172] A. V. Knyazev and K. Neymeyr. “A Geometric Theory for Preconditioned Inverse Iteration III: A Short and Sharp Convergence Estimate for Generalized Eigenvalue Problems”. In: *Linear Algebra and its Applications* 358.1-3 (2003), pp. 95–114. DOI: 10.1016/S0024-3795(01)00461-X (cit. on p. 30).

[173] A. V. Knyazev and K. Neymeyr. “Gradient Flow Approach to Geometric Convergence Analysis of Preconditioned Eigensolvers”. In: *SIAM J. Matrix Anal. & Appl.* 31.2 (2009), pp. 621–628. DOI: 10.1137/080727567 (cit. on p. 30).

[174] M. J. Kochenderfer and T. A. Wheeler. *Algorithms for Optimization*. Cambridge, Massachusetts: The MIT Press, 2019 (cit. on p. 30).

[175] P. Kongkhambut, J. Skulte, L. Mathey, J. G. Cosme, A. Hemmerich, and H. Keßler. “Observation of a Continuous Time Crystal”. In: *Science* 377.6606 (2022), pp. 670–673. DOI: 10.1126/science.abo3382 (cit. on p. 76).

[176] A. Kristof. “Using a Spectral Inference Network to Solve the Time-Independent Schrödinger Equation for a Two-Dimensional Hydrogen Atom”. CES Seminar Thesis. RWTH Aachen University, 2020 (cit. on p. xii).

[177] H. W. Kroto, J. R. Heath, S. C. O’Brien, R. F. Curl, and R. E. Smalley. “C60: Buckminsterfullerene”. In: *Nature* 318.6042 (1985), pp. 162–163. DOI: 10.1038/318162a0 (cit. on p. 76).

[178] A. Kutner and A.-M. Sändig. *Some Applications of Weighted Sobolev Spaces*. Vol. 100. Teubner-Texte zur Mathematik. Wiesbaden: Vieweg+Teubner Verlag, 1987. DOI: 10.1007/978-3-663-11385-0 (cit. on pp. 44, 45).

[179] P. W. Langhoff, J. A. Boatz, R. J. Hinde, and J. A. Sheehy. “Atomic Spectral Methods for Molecular Electronic Structure Calculations”. In: *The Journal of Chemical Physics* 121.19 (2004), pp. 9323–9342. DOI: 10.1063/1.1794634 (cit. on p. 20).

[180] M. G. Larson and F. Bengzon. *The Finite Element Method: Theory, Implementation, and Applications*. Vol. 10. Texts in Computational Science and Engineering. Berlin, Heidelberg: Springer, 2013. DOI: 10.1007/978-3-642-33287-6 (cit. on p. 83).

[181] C. Le Bris. "Computational Chemistry from the Perspective of Numerical Analysis". In: *Acta Numerica* 14 (2005), pp. 363–444. DOI: 10.1017/S0962492904000251 (cit. on pp. 7, 10–17, 19, 20, 22, 23).

[182] S. Lehtola. "A Review on Non-Relativistic, Fully Numerical Electronic Structure Calculations on Atoms and Diatomic Molecules". In: *International Journal of Quantum Chemistry* 119.19 (2019), e25968. DOI: 10.1002/qua.25968 (cit. on p. 22).

[183] S. Lehtola. "Fully Numerical Hartree-Fock and Density Functional Calculations. I. Atoms". In: *International Journal of Quantum Chemistry* 119.19 (2019), e25945. DOI: 10.1002/qua.25945 (cit. on p. 22).

[184] S. Lehtola. "Fully Numerical Hartree-Fock and Density Functional Calculations. II. Diatomic Molecules". In: *International Journal of Quantum Chemistry* 119.19 (2019), e25944. DOI: 10.1002/qua.25944 (cit. on p. 22).

[185] S. Leonardi. "The best constant in weighted Poincaré and Friedrichs inequalities". In: *Rendiconti del Seminario Matematico della Università di Padova* 92 (1994), pp. 195–208 (cit. on p. 52).

[186] A. Levitt. "Convergence of Gradient-Based Algorithms for the Hartree-Fock Equations". In: *ESAIM: Mathematical Modelling and Numerical Analysis* 46.6 (2012), pp. 1321–1336. DOI: 10.1051/m2an/2012008 (cit. on p. 23).

[187] M. Levy. "Universal Variational Functionals of Electron Densities, First-Order Density Matrices, and Natural Spin-Orbitals and Solution of the v-Representability Problem". In: *Proceedings of the National Academy of Sciences* 76.12 (1979), pp. 6062–6065. DOI: 10.1073/pnas.76.12.6062 (cit. on p. 16).

[188] E. H. Lieb. "Density Functionals for Coulomb Systems". In: *International Journal of Quantum Chemistry* 24.3 (1983), pp. 243–277. DOI: 10.1002/qua.560240302 (cit. on pp. 15, 16).

[189] P. L. Lions. "Solutions of Hartree-Fock Equations for Coulomb Systems". In: *Commun.Math. Phys.* 109.1 (1987), pp. 33–97. DOI: 10.1007/BF01205672 (cit. on p. 15).

[190] P.-L. Lions. "On the Schwarz Alternating Method II: Stochastic Interpretation and Order Proprieties". In: *Proceedings of the second international conference on domain decomposition methods for partial differential equations* (1989), pp. 47–70 (cit. on p. 33).

[191] P.-L. Lions. "On the Schwarz Alternating Method III: A Variant for Nonoverlapping Subdomains". In: *Proceedings of the third international symposium on domain decomposition methods for partial differential equations* (1990), pp. 202–223 (cit. on p. 33).

[192] P.-L. Lions. "On the Schwarz Alternating Method. I". In: *First international symposium on domain decomposition methods for partial differential equations* (1988), pp. 1–42 (cit. on pp. 1, 33).

[193] F. Lipparini, B. Stamm, E. Cancès, Y. Maday, and B. Mennucci. "Fast Domain Decomposition Algorithm for Continuum Solvation Models: Energy and First Derivatives". In: *J. Chem. Theory Comput.* 9.8 (2013), pp. 3637–3648. DOI: 10.1021/ct400280b (cit. on p. 77).

[194] J. P. Lowe and K. A. Peterson. *Quantum Chemistry*. 3rd ed. Burlington: Elsevier Academic Press, 2006 (cit. on pp. 7, 8).

[195] C. Lubich. *From Quantum to Classical Molecular Dynamics: Reduced Models and Numerical Analysis*. 1st ed. EMS Press, 2008. DOI: 10.4171/067 (cit. on pp. 7, 9, 10, 13).

[196] S. H. Lui. "Domain Decomposition Methods for Eigenvalue Problems". In: *J. Comput. Appl. Math.* 117.1 (2000), pp. 17–34. DOI: 10.1016/S0377-0427(99)00326-X (cit. on p. 77).

[197] S. H. Lui. "Some Recent Results on Domain Decomposition Methods for Eigenvalue Problems". In: *Proc. Ninth Int. Conf. on Domain Decomposition Methods*. 1996, pp. 426–433 (cit. on p. 77).

[198] T. Mathew. *Domain Decomposition Methods for the Numerical Solution of Partial Differential Equations*. Vol. 61. Springer Science & Business Media, 2008. DOI: 10.1007/978-3-540-77209-5 (cit. on pp. 32, 77).

[199] R. McWeeny. *Methods of Molecular Quantum Mechanics*. Academic Press, 1992 (cit. on pp. 7, 14).

[200] A. Messiah. *Quantum Mechanics*. Vol. 1. Amsterdam: North-Holland, 1961 (cit. on pp. 7, 8).

[201] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi, and N. Marzari. "Two-Dimensional Materials from High-Throughput Computational Exfoliation of Experimentally Known Compounds". In: *Nature Nanotech.* 13.3 (2018), pp. 246–252. DOI: 10.1038/s41565-017-0035-5 (cit. on p. 76).

[202] C. Müller, M. Geratz, C. Heger, and J. Meyer. "Evaluation and Implementation of Schrödinger-Type Eigenvalue Problems in Long Rectangular Domains Using the Finite Element Method". CES Project Thesis. RWTH Aachen University, 2021 (cit. on p. xii).

[203] V. F. Müller. *Quantenmechanik*. München Wien: Oldenbourg, 2000 (cit. on pp. 7, 8).

[204] H. Müntz. "Sur La Solution Des Équations Séculaires et Des Équations Intégrales". In: *C. R. Acad. Sci. Paris* 156 (1913), pp. 860–862 (cit. on p. 26).

[205] H. Müntz. "Solution Directe de l'équation Séculaire et de Quelques Problemes Analogues Transcendants". In: *C. R. Acad. Sci. Paris* 156 (1913), pp. 43–46 (cit. on p. 26).

[206] F. Nataf, H. Xiang, V. Dolean, and N. Spillane. "A Coarse Space Construction Based on Local Dirichlet-to-Neumann Maps". In: *SIAM J. Sci. Comput.* 33.4 (2011), pp. 1623–1642. DOI: 10.1137/100796376 (cit. on p. 37).

[207] K. Neymeyr. "A Geometric Theory for Preconditioned Inverse Iteration I: Extrema of the Rayleigh Quotient". In: *Linear Algebra and its Applications* 322.1-3 (2001), pp. 61–85. DOI: 10.1016/S0024-3795(00)00239-1 (cit. on p. 30).

[208] K. Neymeyr. "A Hierarchy of Preconditioned Eigensolvers for Elliptic Differential Operators". Habilitationsschrift. Universität Tübingen, 2001 (cit. on p. 30).

[209] K. Neymeyr. "A Note on Inverse Iteration". In: *Numerical Linear Algebra with Applications* 12.1 (2005), pp. 1–8. DOI: 10.1002/nla.388 (cit. on p. 30).

[210] K. Neymeyr, E. Ovtchinnikov, and M. Zhou. "Convergence Analysis of Gradient Iterations for the Symmetric Eigenvalue Problem". In: *SIAM J. Matrix Anal. & Appl.* 32.2 (2011), pp. 443–456. DOI: 10.1137/100784928 (cit. on p. 30).

[211] R. A. Nicolaides. "Deflation of Conjugate Gradients with Applications to Boundary Value Problems". In: *SIAM J. Numer. Anal.* 24.2 (1987), pp. 355–365. DOI: 10.1137/0724027 (cit. on p. 36).

[212] J. Nocedal and S. Wright. *Numerical Optimization*. 2nd ed. 2006 Edition. New York: Springer, 2006 (cit. on p. 30).

[213] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, M. I. Katsnelson, I. V. Grigorieva, S. V. Dubonos, and A. A. Firsov. "Two-Dimensional Gas of Massless Dirac Fermions in Graphene". In: *Nature* 438.7065 (2005), pp. 197–200. DOI: 10.1038/nature04233 (cit. on pp. 24, 76).

[214] K. S. Novoselov, A. K. Geim, S. V. Morozov, D. Jiang, Y. Zhang, S. V. Dubonos, I. V. Grigorieva, and A. A. Firsov. "Electric Field Effect in Atomically Thin Carbon Films". In: *Science* 306.5696 (2004), pp. 666–669. DOI: 10.1126/science.1102896 (cit. on pp. 24, 76).

[215] M. M. Pandur. "Preconditioned Gradient Iterations for the Eigenproblem of Definite Matrix Pairs". In: *ETNA* 51 (2019), pp. 331–362. DOI: 10.1553/etna_vol51s331 (cit. on p. 32).

[216] G. Papanicolau, A. Bensoussan, and J.-L. Lions. *Asymptotic Analysis for Periodic Structures*. 1st ed. Vol. 5. North Holland, 1978 (cit. on pp. 51–53, 56).

[217] B. N. Parlett. *The Symmetric Eigenvalue Problem*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1998. DOI: 10.1137/1.9781611971163 (cit. on p. 26).

[218] J. E. Pask and P. A. Sterne. “Finite Element Methods in Ab Initio Electronic Structure Calculations”. In: *Modelling Simul. Mater. Sci. Eng.* 13.3 (2005), R71. DOI: 10.1088/0965-0393/13/3/R01 (cit. on p. 22).

[219] A. K. Pearce, T. R. Wilks, M. C. Arno, and R. K. O’Reilly. “Synthesis and Applications of Anisotropic Nanoparticles with Precisely Defined Dimensions”. In: *Nat. Rev. Chem.* 5.1 (2021), pp. 21–45. DOI: 10.1038/s41570-020-00232-7 (cit. on p. 24).

[220] D. Peterseim, J. Wärnegård, and C. Zimmer. *Super-Localised Wave Function Approximation of Bose-Einstein Condensates*. 2023. arXiv: 2309.11985 [cond-mat] (cit. on p. 77).

[221] E. Polizzi. “Density-Matrix-Based Algorithm for Solving Eigenvalue Problems”. In: *Phys. Rev. B* 79.11 (2009), p. 115112. DOI: 10.1103/PhysRevB.79.115112 (cit. on p. 28).

[222] B. P. Pritchard, D. Altarawy, B. Didier, T. D. Gibson, and T. L. Windus. “New Basis Set Exchange: An Open, Up-to-Date Resource for the Molecular Sciences Community”. In: *J. Chem. Inf. Model.* 59.11 (2019), pp. 4814–4820. DOI: 10.1021/acs.jcim.9b00725 (cit. on p. 20).

[223] A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Numerical Mathematics and Scientific Computation. Oxford, New York: Oxford University Press, 1999 (cit. on p. 32).

[224] A. Reusken and B. Stamm. “Analysis of the Schwarz Domain Decomposition Method for the Conductor-like Screening Continuum Model”. In: *SIAM J. Numer. Anal.* 59.2 (2021), pp. 769–796. DOI: 10.1137/20M1342872 (cit. on pp. 2, 76).

[225] J. Rogel-Salazar. “The Gross-Pitaevskii Equation and Bose-Einstein Condensates”. In: *Eur. J. Phys.* 34.2 (2013), pp. 247–257. DOI: 10.1088/0143-0807/34/2/247 (cit. on p. 19).

[226] C. C. J. Roothaan. “New Developments in Molecular Orbital Theory”. In: *Rev. Mod. Phys.* 23.2 (1951), pp. 69–89. DOI: 10.1103/RevModPhys.23.69 (cit. on pp. 23, 76).

[227] Y. Saad. *Numerical Methods for Large Eigenvalue Problems*. Classics in Applied Mathematics 66. Philadelphia: Society for Industrial and Applied Mathematics, 2011. DOI: 10.1137/1.9781611970739 (cit. on pp. 26, 29, 31, 40).

[228] Y. Saad. *Iterative Methods for Sparse Linear Systems*. Other Titles in Applied Mathematics. Society for Industrial and Applied Mathematics, 2003. DOI: 10.1137/1.9780898718003 (cit. on p. 36).

[229] P. R. Sajanlal, T. S. Sreeprasad, A. K. Samal, and T. Pradeep. “Anisotropic Nanomaterials: Structure, Growth, Assembly, and Functions”. In: *Nano Reviews* 2.1 (2011), p. 5883. DOI: 10.3402/nano.v2i0.5883 (cit. on p. 24).

[230] F. Santosa and M. Vogelius. "First-Order Corrections to the Homogenized Eigenvalues of a Periodic Composite Medium". In: *SIAM J. Appl. Math.* 53.6 (1993), pp. 1636–1668. DOI: 10.1137/0153076 (cit. on p. 56).

[231] H. A. Schwarz. "Ueber Einen Grenzübergang Durch Alternirendes Verfahren". In: *Vierteljahrsschrift der Naturforschenden Gesellschaft* 15 (1870), pp. 272–286 (cit. on pp. 1, 32, 33).

[232] R. Shankar. *Principles of Quantum Mechanics*. New York: Springer US, 1994. DOI: 10.1007/978-1-4757-0576-8 (cit. on p. 7).

[233] J. C. Slater. "The Theory of Complex Spectra". In: *Phys. Rev.* 34.10 (1929), pp. 1293–1322. DOI: 10.1103/PhysRev.34.1293 (cit. on p. 14).

[234] B. F. Smith. "Domain Decomposition Methods for Partial Differential Equations". In: *Parallel Numerical Algorithms*. Ed. by D. E. Keyes, A. Sameh, and V. Venkatakrishnan. ICASE/LaRC Interdisciplinary Series in Science and Engineering. Dordrecht: Springer Netherlands, 1997, pp. 225–243. DOI: 10.1007/978-94-011-5412-3_8 (cit. on p. 32).

[235] N. Spillane, V. Dolean, P. Hauret, F. Nataf, C. Pechstein, and R. Scheichl. "Abstract Robust Coarse Spaces for Systems of PDEs via Generalized Eigenproblems in the Overlaps". In: *Numer. Math.* 126.4 (2014), pp. 741–770. DOI: 10.1007/s00211-013-0576-y (cit. on pp. 36, 37, 77, 84, 92).

[236] N. Spillane. "Robust Domain Decomposition Methods for Symmetric Positive Definite Problems". PhD Thesis. Université Pierre et Marie Curie - Paris VI, 2014 (cit. on p. 77).

[237] N. Spillane, V. Dolean, P. Hauret, F. Nataf, and D. J. Rixen. "Solving Generalized Eigenvalue Problems on the Interfaces to Build a Robust Two-Level FETI Method". In: *Comptes Rendus Mathematique* 351.5-6 (2013), pp. 197–201. DOI: 10.1016/j.crma.2013.03.010 (cit. on p. 37).

[238] B. Stamm. *Mathematical Aspects of Computational Chemistry (Lecture Notes SS20)*. RWTH Aachen University. 2020 (cit. on p. 12).

[239] B. Stamm and L. Theisen. "A Quasi-Optimal Factorization Preconditioner for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains". In: *SIAM J. Numer. Anal.* 60.5 (2022), pp. 2508–2537. DOI: 10.1137/21M1456005 (cit. on pp. iii, xi, xiii, 39).

[240] J. Sun and A. Zhou. *Finite Element Methods for Eigenvalue Problems*. Chapman and Hall/CRC, 2016. DOI: 10.1201/9781315372419 (cit. on pp. 58, 59).

[241] T. A. Suslina. "On Homogenization for a Periodic Elliptic Operator in a Strip". In: *St. Petersburg Math. J.* 16.01 (2004), pp. 237–258. DOI: 10.1090/S1061-0022-04-00849-0 (cit. on p. 42).

[242] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Courier Corporation, 1996 (cit. on pp. 7, 10, 14).

[243] R. A. Tapia, J. E. Dennis, and J. P. Schäfermeyer. "Inverse, Shifted Inverse, and Rayleigh Quotient Iteration as Newton's Method". In: *SIAM Rev.* 60.1 (2018), pp. 3–55. DOI: 10.1137/15M1049956 (cit. on pp. 29, 72).

[244] L. Theisen. "Automated Boundary Layer Mesh Generation for Simulation of Convective Cooling". Bachelor Thesis. RWTH Aachen University, 2018. DOI: 10.18154/RWTH-2023-12261 (cit. on p. 22).

[245] L. Theisen. *P3 Finite Element Basis Functions on 2-Simplex*. 2019. DOI: 10.6084/m9.figshare.9767021.v1 (cit. on p. 21).

[246] L. Theisen. "Simulation of Non-Equilibrium Gas Flows Using the FEniCS Computing Platform". Master Thesis. RWTH Aachen University, 2019. DOI: 10.18154/RWTH-2023-12262 (cit. on pp. 16, 21).

[247] L. Theisen and B. Stamm. *A Scalable Two-Level Domain Decomposition Eigensolver for Periodic Schrödinger Eigenstates in Anisotropically Expanding Domains*. Submitted. 2023. DOI: 10.48550/arXiv.2311.08757. arXiv: 2311.08757 [cs, math] (cit. on pp. iii, xi, xiii, 71).

[248] L. Theisen and B. Stamm. *ddEigenLab.Jl: Domain-Decomposition Eigenvalue Problem Lab (v0.2)*. Zenodo. 2022. DOI: 10.5281/zenodo.6576197 (cit. on pp. xi, 60).

[249] L. Theisen and B. Stamm. *ddEigenLab.Jl: Domain-Decomposition Eigenvalue Problem Lab (v0.3)*. Zenodo. 2023. DOI: 10.5281/zenodo.10121779 (cit. on pp. xi, 94).

[250] L. Theisen and M. Torrilhon. "fenicsR13: A Tensorial Mixed Finite Element Solver for the Linear R13 Equations Using the FEniCS Computing Platform". In: *ACM Trans. Math. Softw.* 47.2 (2021), 17:1–17:29. DOI: 10.1145/3442378 (cit. on pp. xi, 60, 94).

[251] L. Theisen and M. Torrilhon. *fenicsR13: A Tensorial Mixed Finite Element Solver for the Linear R13 Equations Using the FEniCS Computing Platform (v1.4)*. Zenodo. 2020. DOI: 10.5281/zenodo.4172951 (cit. on p. xi).

[252] L. H. Thomas. "The Calculation of Atomic Fields". In: *Math. Proc. Camb. Phil. Soc.* 23.5 (1927), pp. 542–548. DOI: 10.1017/S0305004100011683 (cit. on p. 16).

[253] A. Toselli and O. B. Widlund. *Domain Decomposition Methods–Algorithms and Theory*. Springer Series in Computational Mathematics 34. Berlin: Springer, 2005. DOI: 10.1007/b137868 (cit. on pp. 32, 74, 82, 84).

[254] A. Toth and C. T. Kelley. "Convergence Analysis for Anderson Acceleration". In: *SIAM J. Numer. Anal.* 53.2 (2015), pp. 805–819. DOI: 10.1137/130919398 (cit. on p. 23).

[255] L. N. Trefethen and D. Bau. *Numerical Linear Algebra*. Philadelphia: Society for Industrial and Applied Mathematics, 1997 (cit. on pp. 26, 27, 29).

[256] H. Vandervorst. "Computational Methods for Large Eigenvalue Problems". In: *Handbook of Numerical Analysis* 8 (2002), pp. 3–179. DOI: 10.1016/S1570-8659(02)08003-1 (cit. on p. 26).

[257] M. Vanninathan. "Homogenization of Eigenvalue Problems in Perforated Domains". In: *Proc. Indian Acad. Sci. (Math. Sci.)* 90.3 (1981), pp. 239–271. DOI: 10.1007/BF02838079 (cit. on pp. 42, 44).

[258] K. Varga and J. A. Driscoll. *Computational Nanoscience: Applications for Molecules, Clusters, and Solids*. Cambridge University Press, 2011. DOI: 10.1017/CBO9780511736230 (cit. on p. 68).

[259] R. von Mises and H. Pollaczek-Geiringer. "Praktische Verfahren der Gleichungsauflösung". In: *ZAMM - Zeitschrift für Angewandte Mathematik und Mechanik* 9.2 (1929), pp. 152–164. DOI: 10.1002/zamm.19290090206 (cit. on p. 26).

[260] M. M. Vopson. "Estimation of the Information Contained in the Visible Matter of the Universe". In: *AIP Advances* 11.10 (2021), p. 105317. DOI: 10.1063/5.0064475 (cit. on p. 10).

[261] K. B. Vu, V. V. Vu, H. P. Thi Thu, H. N. Giang, N. M. Tam, and S. T. Ngo. "Conjugated Polymers: A Systematic Investigation of Their Electronic and Geometric Properties Using Density Functional Theory and Semi-Empirical Methods". In: *Synthetic Metals* 246 (2018), pp. 128–136. DOI: 10.1016/j.synthmet.2018.10.007 (cit. on p. 42).

[262] H. Wieland. *Beiträge zur mathematischen Behandlung komplexer Eigenwertprobleme, V. Bestimmung höherer Eigenwerte durch gebrochene Iteration*. Tech. rep. Aerodynamische Versuchsanstalt Göttingen, 1944 (cit. on p. 29).

[263] Wikipedia. *File "Eight Allotropes of Carbon"*. 2006 (cit. on p. 25).

[264] Wikipedia. *File "Nobelpriset i Fysik 2010"*. 2010 (cit. on p. 25).

[265] F. Wilczek. "Quantum Time Crystals". In: *Phys. Rev. Lett.* 109.16 (2012). DOI: 10.1103/PhysRevLett.109.160401 (cit. on p. 76).

[266] J. H. Wilkinson. *The Algebraic Eigenvalue Problem*. Numerical Mathematics and Scientific Computation. Oxford University Press, 1988 (cit. on p. 26).

[267] J. H. Wilkinson and C. Reinsch. *Handbook for Automatic Computation*. Ed. by F. L. Bauer, A. S. Householder, F. W. J. Olver, H. Rutishauser, K. Samelson, and E. Stiefel. Berlin, Heidelberg: Springer, 1971. DOI: 10.1007/978-3-642-86940-2 (cit. on p. 26).

[268] Y. Xi and Y. Saad. "Computing Partial Spectra with Least-Squares Rational Filters". In: *SIAM J. Sci. Comput.* 38.5 (2016), A3020–A3045. DOI: 10.1137/16M1061965 (cit. on p. 28).

[269] F. Xu, Q. Huang, and H. Ma. "A Novel Domain Decomposition Framework for the Ground State Solution of Bose–Einstein Condensates". In: *Comput. Math. Appl.* 80.5 (2020), pp. 1287–1300. DOI: 10.1016/j.camwa.2020.06.014 (cit. on p. 77).

[270] Y. Zhang. "Estimates of Eigenvalues and Eigenfunctions in Elliptic Homogenization with Rapidly Oscillating Potentials". In: *Journal of Differential Equations* 292 (2021), pp. 388–415. DOI: 10.1016/j.jde.2021.05.006 (cit. on p. 42).

[271] X. Zhao, Y. Ando, Y. Liu, M. Jinno, and T. Suzuki. "Carbon Nanowire Made of a Long Linear Carbon Chain Inserted Inside a Multiwalled Carbon Nanotube". In: *Phys. Rev. Lett.* 90.18 (2003), 187401:1–187401:4. DOI: 10.1103/PhysRevLett.90.187401 (cit. on pp. 76, 96).

[272] V. V. Zhikov. "Weighted Sobolev Spaces". In: *Sb. Math.* 189.8 (1998), pp. 1139–1170. DOI: 10.1070/SM1998v189n08ABEH000344 (cit. on p. 52).

[273] M. Zhou, Z. Bai, Y. Cai, and K. Neymeyr. *Convergence Analysis of a Block Preconditioned Steepest Descent Eigensolver with Implicit Deflation*. 2022. arXiv: 2209.03407 [cs, math] (cit. on p. 30).

Curriculum Vitæ

Lambert Theisen was born on the 1st of February 1995 in Bad Ems, Germany. He attended primary school at the *Grundschule Singhofen* and high school at the *Goethe-Gymnasium Bad Ems* and finished the latter with the *Abitur* in March 2014.

In October of the same year, he started his Computational Engineering Science (CES) studies at the RWTH Aachen University. He finished in March 2018 the Bachelor of Science with the thesis *"Automated boundary layer mesh generation for simulation of convective cooling"* under the supervision of Prof. Dr. Manuel Torrilhon. In April 2018, he continued his CES master studies and finished in September 2019 the Master of Science with the thesis *"Simulation of non-equilibrium gas flows using the FEniCS computing platform"* under the supervision of Prof. Dr. Manuel Torrilhon.

In October 2019, he started his Ph.D. studies under the supervision of Prof. Dr. Benjamin Stamm at the MathCCES/ACoM chair at the RWTH Aachen University. He moved with his supervisor in October 2022 to the NMH chair at the University of Stuttgart. During that period, he attended the following national and internal conferences or workshops:

- MOANSI Annual Meeting, online, September 2020.
- Summer school on advanced DD methods, Milano (Italy), November 2021.
- ESCO European Seminar of Computing, Pilsen (Czech Republic), June 2022.
- DD27: 7th International Domain Decomposition Conference, Prague (Czech Republic), July 2022.
- GAMM 92nd Annual Meeting, Aachen (Germany), August 2022.
- SIAM Conference on Computational Science and Engineering (CSE23), Amsterdam (Netherlands), March 2023.
- EMC2 Seminar of the ERC: Extreme-scale Mathematically-based Computational Chemistry, LJLL Paris (France), March 2023.
- GAMM 93rd Annual Meeting, Dresden (Germany), May 2023.
- Workshop on Numerical Analysis of Nonlinear Schrödinger Equations, University of Augsburg (Germany), June 2023.
- 29th Biennial Numerical Analysis Conference, Glasgow (Scotland), June 2023.
- RDM Workshop of the SFB 1481: Sparsity and Singular Structures, RWTH Aachen University (Germany), September 2023.
- MOANSI Annual Meeting, University of Stuttgart (Germany), November 2023.