Polynomial and Rational Approximation for Electronic Structure Calculations

by

Simon Etter

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Mathematics

June 2019

Contents

Acknowledgments

Working towards a PhD is like trying a back flip: until you land it, there is no way of telling whether you are about to pull of the coolest stunt ever or get hurt badly. I consider myself very fortunate to have been allowed to perform my metaphorical back flip under the excellent guidance of Christoph Ortner, who is both an outstanding mathematician and a role model supervisor. Christoph gave me the freedom to pursue all the crazy ideas that came to my mind and simultaneously had the mathematical competence and patience to provide expert guidance on all of them, for which I am very grateful.

It is well known that the success or failure of a back flip is to a large extent determined by the run-up, and PhDs are no different. I would therefore like to express my gratitude towards my mentors at ETH Zurich and my parents for providing me with the best preparation I could have asked for. Furthermore, I would like to thank both the Alexes, Daniel, Geneviève, Julian, Maciej, Ronja and Tom for their friendship and the many discussions, both mathematical and otherwise, that we shared over the past years.

Last but not least, I would like to thank my fiancée Satvin for her constant love and support. Without her, this thesis would not have been finished.

Declarations

I declare that to the best of my knowledge, the material contained in this thesis is original and my own work except where otherwise stated.

This thesis is submitted to the University of Warwick for the degree of Doctor of Philosophy and has not been submitted for a degree at any other university.

Abstract

Atomic-scale simulation of matter has become an important research tool in physics, chemistry, material science and biology as it allows for insights which neither theoretical nor experimental investigation can provide. The most accurate of these simulations are based on the laws of quantum mechanics, in which case the main computational bottleneck becomes the evaluation of functions $f(H)$ of a sparse matrix H (the Hamiltonian).

One way to evaluate such matrix functions is through polynomial and rational approximation, the theory of which is reviewed in Chapter 2 of this thesis. It is well known that rational functions can approximate the relevant functions with much lower degrees than polynomials, but they are more challenging to use in practice since they require fast algorithms for evaluating rational functions $r(H)$ of a matrix argument H. Such an algorithm has recently been proposed in the form of the Pole Expansion and Selected Inversion (PEXSI) scheme, which evaluates $r(H)$ by writing $r(x) = \sum_k \frac{c_k}{x-z_k}$ in partial-fraction-decomposed form and then employing advanced sparse factorisation techniques to evaluate only a small subset of the entries of the resolvents $(H-z)^{-1}$. This scheme scales better than cubically in the matrix dimension, but it is not a linear scaling algorithm in general. We overcome this limitation in Chapter 3 by devising a modified, linear-scaling PEXSI algorithm which exploits that most of the fill-in entries in the triangular factorisations computed by the PEXSI algorithm are negligibly small.

Finally, Chapter 4 presents a novel algorithm for computing electric conductivities which requires evaluating a bivariate matrix function $f(H,H)$. We show that the Chebyshev coefficients $c_{k_1 k_2}$ of the relevant function $f(x_1,x_2)$ concentrate along the diagonal $k_1 \sim k_2$ and that this allows us to approximate $f(x_1,x_2)$ much more efficiently than one would expect based on a straightforward tensor-product extension of the one-dimensional arguments.

Chapter 1

Introduction

Much of modern technology relies on our ability to predict and influence the behaviour of matter at an atomistic scale, e.g. to understand and improve material behaviour in mechanical and civil engineering applications, to develop more effective drugs, or to devise more powerful and efficient computers and batteries. In recent decades, much progress in this direction has been achieved not only through experimentation in the laboratory but also through theoretical investigation and computer simulation. Already in 1966, Robert Mulliken remarked in his Nobel Lecture [Mul66]:

> I would like to emphasize my belief that the era of computing chemists, when hundreds if not thousands of chemists will go to the computing machine instead of the laboratory, for increasingly many facets of chemical information, is already at hand.

Five decades later, chemistry and material science have become some of the largest consumers of computing power both at the Swiss National Supercomputing Centre (CSCS, see Figure 1.1) and on ARCHER, the national supercomputer of the United Kingdom [ARC], and one would presumably find a similar situation in supercomputing centres worldwide. Running simulations on such a large scale requires sophisticated hardware infrastructure and large amounts of energy; hence there is significant interest in developing new algorithms which reduce the amount of computation required to extract macroscopic predictions from the microscopic laws of physics. It is the purpose of this thesis to contribute towards this endeavour. More precisely, this thesis will propose improvements to the simulation techniques of a particular class of electronic structure models introduced in Sections 1.1 and 1.2, as laid out in Section 1.5 after a brief review of some existing simulation techniques in Sections 1.3 and 1.4.

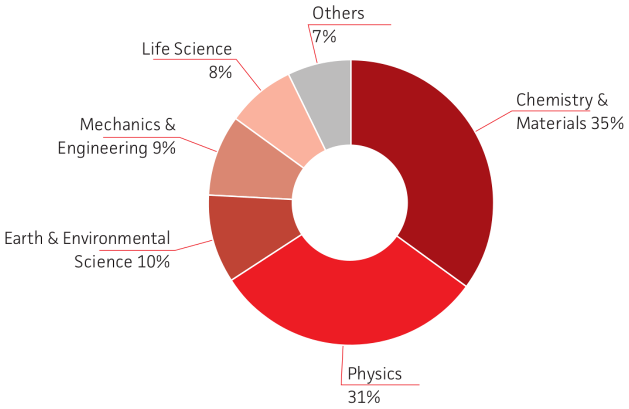


Figure 1.1: Fraction of overall compute time allocated to each research field at the Swiss National Supercomputing Centre in 2017. Figure copied from [CSC17].

1.1 Electronic Structure Models

It is believed that the laws of physics at the atomic level are in principle known but the resulting models are too complicated to allow for computer simulations at the relevant scales which may involve millions or even billions of atoms. In response to this, a ladder of approximate models has been developed where rung by rung reduced accuracy is exchanged for lower computational costs. In this thesis, we will focus on models like Hartree-Fock, Density Functional Theory (DFT) and tight binding which are approximate quantum-mechanical models based on a set of assumptions discussed in the remainder of this section. More in-depth introductions to the topic can be found e.g. in [Kax03, SCS10].

Atoms, electrons and the Born-Oppenheimer approximation. Matter consists of positively charged atomic nuclei and negatively charged electrons. The nuclei are heavy enough that they can be reasonably approximated as discrete point charges which evolve according to classical Newtonian mechanics, while the electrons have to be modeled as quantum-mechanical particles. We therefore represent a system of N nuclei and n electrons through the atomic coordinates $y \in \mathbb{R}^{3N}$ and charges $Z \in \mathbb{R}^N$, and a $3n$-dimensional wave function $\psi(x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}^3$ denoting the electronic coordinates.

Independent-particle approximation. Due to the high-dimensionality, it is impossible to work with a general wave function $\psi(x_1, \ldots, x_n)$ for all but the simplest systems. Instead, we will assume the wave function takes the form of a Slater determinant of single-particle orbitals $\psi_i : \mathbb{R}^3 \to \mathbb{C}$, i.e.

$$\psi(x_1, \ldots, x_n) = \frac{1}{\sqrt{n!}} \sum_{\pi} \operatorname{sign}(\pi) \prod_{i=1}^{n} \psi_i(x_{\pi_i}) \tag{1.1}$$

where the sum runs over all permutations π on $\{1, \ldots, n\}$. This ansatz may be interpreted either as an approximate solution to the exact model (Hartree-Fock), as the exact solution to an approximate model (Kohn-Sham Density Functional Theory (DFT)), or simply as an empirical model fitted to reproduce experimental data (tight binding). The Slater determinant is fully specified once the orbitals ψ_i are known; hence the information contained in (1.1) can be equivalently represented in the single-particle density matrix

$$\begin{aligned} \Gamma(x, x') &:= \int \ldots \int \psi(x, x_2, \ldots, x_n)\, \overline{\psi(x', x_2, \ldots, x_n)}\, dx_2 \ldots dx_n \\ &= \sum_{i=1}^{n} \psi_i(x)\, \overline{\psi_i(x')}. \end{aligned} \tag{1.2}$$

Note that to derive the expression on the second line, we must assume the orbitals ψ_i to be orthogonal. This will always be the case since the orbitals ψ_i are eigenvectors of a Hermitian operator as we shall see next.

Eigenvalue equations. In all three of the aforementioned independent-particle approximations, the orbitals ψ_i are determined as the eigenfunctions of some Hamiltonian operator H given by

$$(H\psi)(x) = -\Delta\psi(x) + V(y, Z, x)\, \psi(x),$$

where the potential $V(y, Z, x)$ represents the interaction between the electrons and the atoms, and the interactions among the electrons themselves. The eigenvalues ε_i associated with the eigenfunction ψ_i are interpreted as the energy of the electron occupying the orbital ψ_i. The density matrix Γ formed by the n orbitals ψ_i of lowest energies ε_i is known as the ground state, while any other combination of orbitals is referred to as an excited state.

Self-consistency. The electron-electron interaction part of the potential $V(y, Z, x)$ generally depends on the electronic density $\rho(x)$ which in turn is a function of the orbitals $\psi_i(x)$; hence the eigenvalue equation $H\psi_i = \varepsilon_i\, \psi_i$ is generally a non-linear one. This non-linearity is usually tackled by means of the self-consistent field iteration which solves the linearised eigenvalue problem repeatedly until a fixed point is reached.

Fermi-Dirac distribution. A density matrix of the form (1.2) is known as a

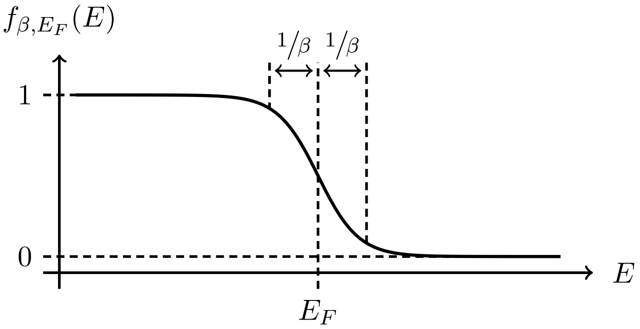


Figure 1.2: The Fermi-Dirac function f_{β,E_F}.

pure quantum-mechanical state because it corresponds to a single wave function $\psi(x_1,\ldots,x_n)$. Such pure states occur rarely in nature since the interaction with an environment at finite temperature T quickly causes the density matrix to relax into a superposition of pure states of the form

$$\Gamma(x,x') = \sum_i f_{\beta,E_F}(\varepsilon_i)\,\psi_i(x)\,\overline{\psi_i(x')}, \tag{1.3}$$

where the Fermi-Dirac function

$$f_{\beta,E_F}(E) := \frac{1}{1+\exp\bigl(\beta\,(E-E_F)\bigr)} \tag{1.4}$$

is given by a step function centred at the Fermi energy E_F and smeared according to the inverse temperature $\beta := \frac{1}{T}$, see Figure 1.2. In physical terms, the finite-temperature density matrix (1.3) describes a system of electronic states ψ_i coupled to an infinite pool of electrons at energy E_F. At zero temperature or equivalently $\beta = \infty$, the electrons flow from the pool into the system until every state below the Fermi energy E_F is fully occupied, and the electrons in states above the Fermi energy get drained into the pool and hence remain empty. The Fermi-Dirac function $f_{\beta,E_F}(E)$, which describes the occupancy of a state at energy E, hence takes the form of a sharp step function and the density matrix (1.3) equals the pure density matrix from (1.2) in this case. At finite temperatures $\beta < \infty$, on the other hand, thermal fluctuations excite some electrons in states slightly below the Fermi energy into states slightly above the Fermi energy, which causes the smearing shown in Figure 1.2.

Metals and insulators. In the limit of large systems $N \to \infty$, the eigenvalues ε_i

of H converge to the the spectral measure

$$\mu(E) := \lim_{N\to\infty} \sum_i \delta(E - E_i),$$

and materials may be classified depending on whether there exists an interval $(\varepsilon_-, \varepsilon_+) \subset \mathbb{R}$ containing the Fermi energy E_F such that

$$\mu\big((\varepsilon_-, \varepsilon_+)\big) = 0. \tag{1.5}$$

If (1.5) holds, the material is called an insulator with band gap $\delta E := \varepsilon_+ - \varepsilon_-$, while if not the material is called a metal. Insulators with a small band gap δE are also known as semiconductors.

1.2 Quantities of Interest

The following is a selection of physical observables which can be computed using the above electronic structure model. For all but the last observable, we provide an expression both in terms of a sum over electronic states ψ_i which allows for convenient physical interpretation, and in terms of some function $f(H)$ of the Hamiltonian H which will be important for the algorithmic developments in this thesis. The equivalence between the two expressions follows from the identity $H = \sum_i \varepsilon_i \, |\psi_i\rangle\langle\psi_i|$.

- Number of electrons [Goe99, eq. (14)],

  $$n = \sum_i f_{\beta,E_F}(\varepsilon_i) = \mathrm{Tr}\big(f_{\beta,E_F}(H)\big), \tag{1.6}$$

  which allows us to determine the Fermi energy E_F in applications where the number of electrons n rather than the Fermi energy E_F is prescribed.

- Electronic density [Goe99, eq. (17)],

  $$\rho(x) := \sum_i f_{\beta,E_F}(\varepsilon_i) \, |\psi_i(x)|^2 = \mathrm{diag}\big(f_{\beta,E_F}(H)\big), \tag{1.7}$$

  which provides insight into chemical bonding and is required in self-consistent field iterations.

- Total electronic energy [Goe99, eq. (15)] and force on atom $I \in \{1, \dots, N\}$

[Goe99, eq. (56)],

$$
\begin{aligned}
E_{tot} &:= \sum_i f_{\beta,E_F}(\varepsilon_i)\,\varepsilon_i = \operatorname{Tr}\big(H\, f_{\beta,E_F}(H)\big), \\
F_I &:= -\frac{\partial E_{tot}}{\partial y_I} = \sum_i f_{\beta,E_F}(y_I)\,\langle\psi_i|\tfrac{\partial H}{\partial y_I}|\psi_i\rangle, = \operatorname{Tr}\Big(f_{\beta,E_F}(H)\,\tfrac{\partial H}{\partial y_I}\Big),
\end{aligned}
$$

which allow us to find equilibrium configurations of the atoms and perform molecular dynamics simulations.

- Conductivity tensor [Kax03, eq. (5.45)]

$$
\sigma_{a,b} = \sum_{i_1,i_2} F_\zeta(\varepsilon_{i_1},\varepsilon_{i_2})\,\langle\psi_{i_1}|M_a|\psi_{i_2}\rangle\,\langle\psi_{i_2}|M_b|\psi_{i_1}\rangle, \qquad a,b \in \{1,2,3\}, \tag{1.8}
$$

which expresses the linear relationship between the electric field $\vec{E}$ and the induced current $\vec{J}$, i.e. $\vec{J} = \sigma\,\vec{E}$. The conductivity function $F_\zeta(E_1,E_2)$ is given by

$$
F_\zeta(E_1,E_2) = \frac{f_{\beta,E_F}(E_1) - f_{\beta,E_F}(E_2)}{E_1 - E_2}\,\frac{1}{E_1 - E_2 + \omega + i\eta} \tag{1.9}
$$

and depends on β and E_F discussed above, ω (the oscillation frequency of the electric field) and η (the inverse relaxation time, which is an empirical parameter measuring the mobility of electrons in a given material). For notational convenience, we collect these four parameters into a single variable $\zeta = (\beta, E_F, \omega, \eta)$. The velocity operators M_a are given by $M_a = i\,\big(X_a\,H - H\,X_a\big)$ with $(X_a\psi)(x) := x_a\,\psi(x)$. In this thesis, the symbol i is used to denote both an index $i \in \mathbb{Z}$ and the imaginary unit $i = \sqrt{-1}$. Context will clarify the intended meaning.

1.3 Electronic Structure Algorithms

This thesis focuses on the algorithmic aspects of electronic structure models, and hence we will always be working with discretised Hamiltonians given as finite matrices $H \in \mathbb{R}^{m\times m}$ with entries $H(i,j)$ bounded independently of N. The spectra $\mathcal{E} = \big\{\varepsilon_i \mid i \in \{1,\dots,m\}\big\}$ of such matrices are discrete but the eigenvalues typically cluster in a small number of finite intervals $\mathcal{E}_k \subset \mathbb{R}$ such that we will assume $\mathcal{E}$ to be the union of these intervals for most of this thesis. Without loss of generality, we will further assume the Hamiltonian H to be shifted and scaled such that $\mathcal{E} \subseteq [-1,1]$. The only restriction imposed on the discretisation is that the resulting matrices are required to be sparse, i.e. we assume the discretisation is performed using localised basis functions like atomic orbitals or finite elements, and we exclude spatially extended basis sets like plane waves. Furthermore, we will restrict our attention to the linearised eigenvalue problem which as we have indicated above serves as an important building block of the self-consistent field iteration for tackling the nonlinear problem.

Under these circumstances, the quantities of interest from Section 1.2 are in principle straightforward to evaluate: compute the eigenpairs ψ_i, ε_i of H using e.g. the QR algorithm, and insert these quantities into the respective formulae. This approach is known as the *diagonalisation algorithm* due to its reliance on the eigenvalue decomposition, and while conceptually simple it suffers from the major drawback that diagonalising H scales cubically in the matrix size m which in turn typically grows linearly with the number of atoms N. This cubic scaling effectively limits the diagonalisation algorithm to systems of at most one or two thousand atoms [Hin17, OTBM16, MRG⁺15].

Over the past four decades, several alternative algorithms have been proposed which aim to extend the reach of electronic structure models beyond this "cubic scaling wall" by reducing the simulation cost to $\mathcal{O}(N)$. All of these linear scaling algorithms are based on the observations that (1) the quantities of interest from Section 1.2 can be computed easily once the density matrix $f_{\beta,E_F}(H)$ is available, and (2) for insulators and metals at finite temperature $\beta < \infty$, the density matrix $f_{\beta,E_F}(H)$ is localised or near-sighted [Koh96, BBR13], i.e. we have that

$$|f_{\beta,E_F}(H)(i,j)| \leq C \, \exp\bigl(-\gamma \, d(i,j)\bigr) \tag{1.10}$$

for some constants C, γ and some notion of distance $d(i,j)$ independent of the system size N. The density matrix $f_{\beta,E_F}(H)$ has hence only $\mathcal{O}(m)$ significant entries, which raises hope that an approximation $\tilde{\Gamma} \approx f_{\beta,E_F}(H)$ can be computed with only $\mathcal{O}(m)$ runtime as well. This is indeed possible, and in the remainder of this section we introduce three frequently used strategies for computing $\tilde{\Gamma}$ efficiently. We refer to the review articles [Goe99, BM12] for more details regarding the algorithms presented below and other linear scaling algorithms.

Domain decomposition. The same techniques which prove the localisation (1.10) can also be used to show that the entries $f_{\beta,E_F}(H)(i,j)$ of the density matrix depend exponentially weakly on the Hamiltonian entries $H(i',j')$ "far away" from (i,j), i.e.

we have that

$$\left| \frac{\partial f_{\beta,E_F}(H)(i,j)}{\partial H(i',j')} \right| \leq C \, \exp\Big(-\gamma \left(d(i,i') + d(j',j) \right) \Big). \tag{1.11}$$

This suggests that we evaluate a single entry $f_{\beta,E_F}(H)(i,j)$ of the density matrix by truncating H to some buffer region $B \subset \{1,\dots,m\}$ around (i,j) and approximating

$$f_{\beta,E_F}(H)(i,j) \approx f_{\beta,E_F}\big(H(B,B)\big)(i,j), \tag{1.12}$$

where we note that the error in (1.12) decays exponentially in the buffer size B and independently of the system size N as a consequence of (1.11). We can thus evaluate all the $\mathcal{O}(m)$ significant entries of the density matrix in only $\mathcal{O}(m)$ runtime even if the diagonalisation algorithm is used to perform (1.12) since the buffer size B remains bounded for growing system sizes N.

The domain decomposition method as presented above is rather inefficient in practice due to the large overlap between buffer regions for different entries (i,j), which causes closely related computations in the overlap regions to be performed repeatedly. However, this method becomes highly effective if the density matrix exhibits some sort of regularity such that it can be reconstructed from only few sampled entries $f_{\beta,E_F}(H)(i,j)$, see e.g. [MLO17] and Chapter 4. Furthermore, the domain decomposition method has been used as a building block in multiscale methods [CO16], and it served as a theoretical tool for thermodynamic limits arguments in [MLO17].

Function approximation. We have seen in Section 1.2 that the quantities of interest q can be equivalently defined either in terms of sums over eigenstates ψ_i or as (weighted) traces of functions of the Hamiltonian H,

$$q = \sum_i f(\varepsilon_i) \, \langle \psi_i | M | \psi_i \rangle = \operatorname{Tr}\big(M \, f(H)\big), \tag{1.13}$$

the only exceptions being the electronic density (1.7), to which very similar arguments apply, and the conductivity (1.8) which will be discussed in Chapter 4. Starting from the last expression in (1.13), we note that the diagonalisation algorithm may be interpreted as just a particular method for evaluating the matrix function $f(H)$, and if we formulate the problem this way a possible solution to the cubic scaling problem suggests itself: instead of evaluating $f(H)$ exactly using the eigendecomposition of H, we may evaluate $f(H)$ approximately by first determining a polynomial approximation $p(E) \approx f(E)$ and then replacing $f(H)$ by $p(H)$ in (1.13). The approximate quantity $\tilde q := \operatorname{Tr}\big(M \, p(H)\big)$ computed in this way satisfies the error bound

$$\begin{aligned} |q - \tilde{q}| &= \Big| \operatorname{Tr}\Big(M \, \big(f(H) - p(H) \big) \Big) \Big| \\ &\leq \|\mathrm{vec}(M)\|_1 \, \|f(H) - p(H)\|_{\mathrm{nz}(M)} \\ &= \mathcal{O}(m) \, \big\| f - p \big\|_{\mathcal{E}} \end{aligned} \tag{1.14}$$

where $\|\mathrm{vec}(M)\|_1 := \sum_{i,j=1}^m |M(i,j)| = \mathcal{O}(m)$ due to sparsity, $\mathrm{nz}(M) := \big\{(i,j) \in \{1,\ldots,m\}^2 \mid M(i,j) \neq 0\big\}$, $\|A\|_{\mathcal{I}} := \max_{(i,j)\in\mathcal{I}} |A(i,j)|$ and $\|f\|_{\mathcal{E}} := \sup_{E\in\mathcal{E}} |f(E)|$. The $\mathcal{O}(m)$-factor in (1.14) suggests that keeping the approximation p fixed for growing system sizes N results in a constant error per atom which is usually acceptable in applications. The only part which scales with the system size N is thus the matrix size m, and since H is sparse it follows that $p(H)$ can be evaluated in $\mathcal{O}(m)$ cost. Notable examples of this class of linear-scaling algorithms include the Fermi operator expansion [GC94, GT95] and the Kernel polynomial method [SRVK96, VKS96].

Minimisation methods. The ground state density matrix from (1.2) is the unique local minimiser of

$$\mathcal{F}(\Gamma) := \operatorname{Tr}\Big(\big(3\Gamma^2 - 2\Gamma^3\big) \, (H - \mu I) \Big) \tag{1.15}$$

restricted to the set of Hermitian matrices; thus yet another way to achieve linear scaling is to minimise $\mathcal{F}(\Gamma)$ or related functionals over the space of band limited matrices. Alternatively, the ground state orbitals ψ_i may be determined with linear-scaling cost by solving a minimisation problem in terms of localised trial orbitals $\tilde{\psi}_i$, i.e. orbitals $\tilde{\psi}_i$ centred at points $x_i \in \mathbb{R}^3$ such that $\tilde{\psi}_i(x - x_i) = 0$ if $|x - x_i|$ is larger than some cut-off radius r. This thesis will not discuss minimisation methods, but we nevertheless decided to mention them here since most electronic structure codes in use today are based on some form of minimisation, e.g. CONQUEST [BM10] (density matrix minimisation), SIESTA [SAG$^+$02] and QUICKSTEP [VKM$^+$05] (orbital minimisation), and ONETEP [SHMP05] and BigDFT [MRG$^+$15] (combination of density matrix and orbital minimisation).

1.4 Pole Expansion and Selected Inversion

A shared drawback of all the linear-scaling algorithms mentioned above is that their performance rapidly degrades if the localisation rate γ from (1.10) decreases. A new algorithm in the class of function approximation methods, called the *pole expansion and selected inversion (PEXSI)* algorithm, has recently been introduced in [LCYH13] which overcomes this limitation to a large extent. This method proceeds by (1) approximating the Fermi-Dirac function by a rational function $r_{\tilde{n}}(E)$ in pole-expanded form,

$$r_{\tilde n}(E) := \sum_{k=1}^{\tilde n} \frac{c_k}{E - z_k} \approx f_{\beta,E_F}(E), \tag{1.16}$$

and (2) evaluating only the entries $(H-z)^{-1}(i,j)$ with (i,j) in the nonzero structure

$$\mathrm{nz}(H) := \big\{(i,j) \in \{1,\dots,m\}^2 \mid H(i,j) \neq 0\big\} \tag{1.17}$$

of H. This is sufficient to evaluate the quantities of interest listed in Section 1.2 (other than the conductivity, which will be discussed in Chapter 4), since closer inspection of the formulae in Section 1.2 reveals that in fact only the entries

$$f_{\beta,E_F}(H)(i,j) \approx \sum_{k=1}^{\tilde n} c_k\,(H - z_k)^{-1}(i,j)$$

with $(i,j) \in \mathrm{nz}(H)$ are needed to this end. The PEXSI algorithm thus decomposes the problem of simulating electronic structure problems into the two subproblems of (1) finding a rational approximation of the form (1.16) with as few poles $\tilde n$ as possible and (2) evaluating $(H-z)^{-1}(i,j)$ with $(i,j) \in \mathrm{nz}(H)$ as fast as possible.

The first subproblem has been addressed in [LLYE09], where an exponentially convergent rational approximation scheme of the form (1.16) was constructed and the rate of decay γ in the error bound

$$\|r_{\tilde n} - f_{\beta,E_F}\|_{\mathcal{E}} \leq C\,\exp\!\big(-\gamma\,\tilde n\big).$$

was shown to be lower-bounded by[1] $\gamma \gtrsim |\log(\delta E)|^{-1} + \log(\beta)^{-1}$ where we recall from Section 1.1 that β denotes the inverse temperature and δE denotes the band gap. This logarithmic dependence of the convergence rate on the model parameters in the case of rational approximation should be compared against the algebraic dependence $\gamma \sim \delta E + \beta^{-1}$ in the polynomial case (see Theorem 2.3.12), which demonstrates that rational functions perform much better than polynomials at approximating the small-band-gap, low-temperature Fermi-Dirac function. An iterative scheme to find optimal rational approximations has later been proposed in [Mou16] and was found to produce approximations of about the same accuracy as the scheme from [LLYE09] with four times fewer poles.

Regarding the second PEXSI subproblem, it has been noted in [LLY$^+$09, LYM$^+$11] that the entries $(H-z)^{-1}(i,j)$ with $(i,j) \in \mathrm{nz}(H)$ can be evaluated efficiently using the selected inversion algorithm from [ET75] which consists of two steps.

[1] The notation $\sim$, $\lesssim$ and $\gtrsim$ is defined in Appendix A.2.

	Runtime	Memory	Example system
$d = 1$	$\mathcal{O}(m)$	$\mathcal{O}(m)$	Nanotubes
$d = 2$	$\mathcal{O}(m^{3/2})$	$\mathcal{O}(m \log(m))$	Monolayers
$d = 3$	$\mathcal{O}(m^2)$	$\mathcal{O}(m^{4/3})$	Bulk solids

Table 1.1: Compute time and memory costs of the selected inversion algorithm depending on the effective dimension d of the atomic system. Citations and a discussion of this result can be found in Subsection 3.1.2.

- Compute the factorisation $H - z = LDL^T$ where $L \in \mathbb{C}^{m \times m}$ is lower-triangular with unit diagonal and $D \in \mathbb{C}^{m \times m}$ is diagonal.

- Evaluate the entries $(H - z)^{-1}(i, j)$ with (i, j) in the set

  $$\mathrm{fnz}(H) := \big\{(i,j) \in \{1, \ldots, m\}^2 \mid L(i,j) \neq 0 \vee L(j,i) \neq 0\big\}. \tag{1.18}$$

  This can be done using only the factors L, D and entries $(H - z)^{-1}(i, j)$ with $(i, j) \in \mathrm{fnz}(H)$ computed recursively.

Both of these steps incur the same asymptotic costs listed in Table 1.1. We will use the term *selected inversion algorithm* to refer to the combination of the above steps and *selected inversion step* or *subalgorithm* to refer to just the second step.

We infer from the costs listed in Table 1.1 that the PEXSI algorithm is only a reduced-order method but not a linear-scaling one, i.e. its scales better than the diagonalisation algorithm but its cost is not $\mathcal{O}(N)$ in dimensions $d > 1$. Despite this, the PEXSI method combines several features which make it a viable alternative to the diagonalisation method and linear-scaling algorithms.

- Excellent parallel scaling up to 100,000 processors has been demonstrated in [LGHY14, YCG+18].

- The method is virtually a black-box electronic structure solver since it involves only a single approximation parameter (the number of poles n) whose impact on the runtime and accuracy can be quantified before running any large-scale simulations.

1.5 Contributions

The first part of this thesis will propose a modification to the PEXSI method which reduces the runtime and memory complexities reported in Table 1.1 to $\mathcal{O}(m)$ for all dimensions. Like all linear-scaling methods, our modification is based on the localisation phenomenon described in (1.10), and we will see in Chapter 3 that this phenomenon may be understood as a consequence of the convergence of polynomial approximation to the Fermi-Dirac function. Chapter 2 therefore reviews polynomial approximation in one dimension, and it also discusses the extension of this theory to rational approximation in order to provide some context for the PEXSI method. Chapter 3 will then show that the triangular factorisation computed by the selected inversion algorithm exhibits a form of localisation similar to that of the density matrix described in (1.10), and it will present and analyse the aforementioned linear-scaling modification to the PEXSI method.

In the second part of this thesis, we will present in Chapter 4 a novel algorithm for evaluating the conductivity (1.8) which requires a bivariate polynomial or rational approximation $p(E_1, E_2) \approx F_\zeta(E_1, E_2)$ to the conductivity function $F_\zeta(E_1, E_2)$ defined in (1.9). We will see that the Chebyshev coefficients of this function exhibit a particular asymptotic decay which allows us to significantly reduce the costs of the aforementioned algorithm in the regime of large inverse temperatures β and small inverse relaxation times η.

Chapter 2

Approximation of the Fermi-Dirac Function

We have seen in Section 1.4 that the PEXSI scheme requires a rational function $r(x)$ which approximates the Fermi-Dirac function f_{β,E_F} from (1.4) on the spectrum $\mathcal{E}$ of the Hamiltonian, and we will see in Chapter 3 that the analogous polynomial approximation problem is related to the localisation phenomenon described in Section 1.3. This chapter discusses both of these problems in a unified framework.

As a composition of analytic functions, the Fermi-Dirac function

$$f_{\beta,E_F}(E) := \frac{1}{1+\exp\big(\beta\,(E-E_F)\big)}$$

is analytic everywhere except on the set

$$\mathcal{S}_{\beta,E_F} := \big\{E_F + \tfrac{\pi i k}{\beta} \mid k \text{ odd}\big\} \tag{2.1}$$

where the denominator becomes zero. This function is thus in particular analytic on the domain of approximation $\mathcal{E} \subset \mathbb{R}$, and the above problems may be formulated in abstract terms as follows.

Problem 2.0.1 *Given a closed domain of approximation $\mathcal{E} \subset \mathbb{C}$, a closed set of singularities $\mathcal{S} \subset \mathbb{C} \setminus \mathcal{E}$, and an analytic and single-valued function $f : \mathbb{C} \setminus \mathcal{S} \to \mathbb{C}$, find a rational function $r(x)$ of numerator degree m and denominator degree n such that the supremum norm $\|f - r\|_{\mathcal{E}}$ of the error on $\mathcal{E}$ is small. We will always assume that $\mathcal{E}$ and $\mathcal{S}$ have nonzero logarithmic capacity as defined in Definition* 2.1.5.

We deliberately formulated Problem 2.0.1 in terms of a vague "smallness" criterion since in the following we will be concerned with constructing and analysing a concrete rational approximation scheme rather than finding the best possible approximation. Furthermore, we would like to emphasise that for consistency with the literature on rational approximation (see e.g. [Tre13]), the variables m and n in this chapter refer, respectively, to the numerator and denominator degrees of a rational function $r(x)$ rather than to the number of degrees of freedom and number of electrons as in Chapter 1. Finally, we remark that Problem 2.0.1 includes polynomial approximation as the special case $n = 0$, and the discussion in Sections 2.1 and 2.2 applies equally to both the polynomial as well as the truly rational case. Only from Section 2.3 onward will we start to distinguish these two cases.

The theory associated with Problem 2.0.1 is closely related to the field of logarithmic potential theory, which studies the electrostatic potential induced by charged conductors in the complex plane. Section 2.1 will give a brief introduction to logarithmic potential theory and its connection with approximation theory, and we will see in Theorem 2.1.10 that Problem 2.0.1 can be tackled by determining the equilibrium distribution of two sets of electric charges restricted to $\mathcal{E}$ and $\mathcal{S}$, respectively. Section 2.2 will then discuss a few technical tools required to determine this equilibrium distribution, and Sections 2.3 and 2.4 will demonstrate the application of these tools to polynomial and rational approximation of the Fermi-Dirac function, respectively. Finally, Sections 2.5 and 2.6 will compare the rational interpolation scheme proposed in this chapter against rational approximation via contour quadrature and Zolotarev's best rational approximations to the sign function, respectively.

Our discussion of polynomial approximation is the result of applying the standard theory as presented e.g. in [Tre13, Saf10] in the context of electronic structure theory. Rational approximations to the Fermi-Dirac function have been constructed previously in [LLYE09, Mou16], and we will see in Sections 2.4 and 2.5 that our rational approximation scheme outperforms the one from [LLYE09] by a factor of two, but underperforms the optimal rational approximations determined in [Mou16] by a factor of two. Hence, the mathematical results presented in this chapter are either not new or of little practical relevance, but we believe that there is some benefit in presenting the various results from the literature in a single and coherent framework.

Definition 2.0.2 *This chapter will use the following notation.*

- $\|f\|_S := \sup_{x \in S} |f(x)|$ *denotes the supremum norm of* $f(x)$ *on a set* $S \subset \mathbb{C}$.
- $\mathcal{P}_n$ *denotes the space of polynomials of degree* $\leq n$.
- $\mathcal{R}_{mn}$ *denotes the space of rational functions of numerator degree* $\leq m$ *and denominator degree* $\leq n$.

- *$\delta(x)$ denotes the Dirac delta measure defined by $\int f(x)\,d\delta(x) = f(0)$.*
- *$\int_{\partial\Omega} f(x)\,dx$ denotes the contour integral along $\partial\Omega$ taken in counterclockwise direction relative to the interior of Ω.*
- *In the context of the previous item, $\partial\gamma$ for a curve $\gamma \subset \mathbb{C}$ denotes the counterclockwise contour around a domain of infinitesimal width. For example, we set*
$$\partial[-1,1] = \big([-1,1] + 0i\big) \cup \big([-1,1] - 0i\big),$$
where the signed zero in the imaginary part indicates which branch to evaluate for a function with branch cut along $[-1,1]$.

2.1 Rational Interpolation and Logarithmic Potential Theory

The connection between logarithmic potentials and approximation theory is most easily seen at the example of rational interpolation, which we introduce through the following theorem.

Theorem 2.1.1 ([Wal56, Theorem 8.1]) *Given a function $f : \mathbb{C} \to \mathbb{C}$, a set of distinct interpolation points $X = \{x_0, \dots, x_m\} \subset \mathbb{C}$ and a set of distinct poles $Y = \{y_1, \dots, y_n\} \subset \mathbb{C}$ such that $X \cap Y = \emptyset$, there exists a unique rational function $r \in \mathcal{R}_{mn}$ of the form*

$$r(x) = \frac{p(x)}{\ell_Y(x)} \qquad \text{with} \qquad p(x) \in \mathcal{P}_n \qquad \text{and} \qquad \ell_Y(x) := \prod_{\ell=1}^{n} (x - y_k)$$

such that $r(x_k) = f(x_k)$ for $k \in \{0, \dots, m\}$.

Given X, Y and f as in Theorem 2.1.1, the rational interpolant is easily evaluated using the barycentric interpolation formula, see [Tre13, §5], and the following theorem allows us to estimate the resulting error.

Theorem 2.1.2 (Hermite interpolation formula, [Wal56, Theorem 8.2]) *Let $\mathcal{E}$, $\mathcal{S}$ and $f(x)$ be as in Problem 2.0.1. Then, the rational interpolant $r(x)$ to $f(x)$ at points $X = \{x_0, \dots, x_m\} \subset \mathcal{E}$ with poles $Y = \{y_1, \dots, y_n\} \subset \mathbb{C} \setminus \mathcal{E}$ satisfies for all $x \in \mathcal{E}$*

$$f(x) - r(x) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{\ell_X(x)}{\ell_Y(x)} \, \frac{\ell_Y(t)}{\ell_X(t)} \, \frac{f(t)}{t - x} \, dt \tag{2.2}$$

where $\ell_Z(x) := \prod_{z_k \in Z} (x - z_k)$ for any discrete set $Z \subset \mathbb{C}$ and $\Omega \subset \mathbb{C}$ denotes an open set such that $\mathcal{E} \subset \Omega$ and closure$(\Omega) \subset \mathbb{C} \setminus \mathcal{S}$.

Remark 2.1.3 The set Ω must be introduced in Theorem 2.1.2 since $f(x)$ is typically unbounded on $\partial\mathcal{S}$ and thus the integral (2.2) would be undefined if we set $\Omega := \mathbb{C}\setminus\mathcal{S}$. However, the conclusions which we will draw from Theorem 2.1.2 will be the sharpest if we take the limit $\Omega \to \mathbb{C}\setminus\mathcal{S}$, and in order to simplify the exposition we already anticipate this limit by writing $\mathbb{C}\setminus\mathcal{S}$ instead of Ω in the following. We will return to this issue in Remark 2.1.11.

Replacing $\Omega \to \mathbb{C}\setminus\mathcal{S}$ as discussed in Remark 2.1.3, Theorem 2.1.2 implies the bound

$$\|f(x) - r(x)\|_{\mathcal{E}} \le C\,\|\ell_X/\ell_Y\|_{\mathcal{E}}\,\|\ell_Y/\ell_X\|_{\partial\mathcal{S}} \tag{2.3}$$

where

$$C = \frac{1}{2\pi}\int_{\partial\mathcal{S}} \frac{|f(t)|}{|t-x|}\,|dt|;$$

hence in order to make the approximation error $|f(x) - p(x)|$ small on $\mathcal{E}$, the interpolation points X and poles Y should be chosen such that the ratio ℓ_X/ℓ_Y becomes uniformly small on $\mathcal{E}$ but large on $\partial\mathcal{S}$. We note that

$$U_{X,Y}(x) := \log\frac{|\ell_X(x)|}{|\ell_Y(x)|} = \sum_{k=0}^{m}\log|x - x_k| - \sum_{\ell=1}^{n}\log|x - y_\ell|$$

is the electrostatic potential of a system with charges -1 at each point $x_k \in X$ and charges $+1$ at each point $y_\ell \in Y$, and we will see in Theorem 2.1.8 below that in the limit $m, n \to \infty$, the configurations X, Y which minimise the potential energy

$$I_{X,Y} := -\sum_{k=0}^{m} U_{X\setminus\{x_k\},Y}(x_k) + \sum_{\ell=1}^{n} U_{X,Y\setminus\{y_\ell\}}(y_\ell)$$

simultaneously minimise the bound (2.3). We therefore discuss next a few key results regarding the limiting distributions of the point sets X, Y and their associated potentials. Textbooks and manuscripts on the material presented here can be found e.g. in [Saf10, ST97, Ran95].

Theorem 2.1.4 *Let $\mathcal{E}, \mathcal{S}$ and m, n be as in Problem* 2.0.1*, and let $\mathcal{E}$ be bounded if $m \ge n$ and $\mathcal{S}$ be bounded if $n \ge m$. Then, there exists a unique Borel measure $\mu_{\mathcal{E},\mathcal{S},{}^{n}/_{m}}$, called the equilibrium measure, which minimises the energy*

$$I(\mu) := -\iint \log|x - t|\,d\mu(t)\,d\mu(x) \tag{2.4}$$

over all Borel measures μ supported on $\mathcal{E} \cup \mathcal{S}$ such that μ is positive on $\mathcal{E}$, negative on $\mathcal{S}$, and we have that

$$\mu(\mathcal{E}) = 1, \qquad \mu(\mathcal{S}) = -\tfrac{n}{m}.$$

This equilibrium measure is supported on the boundary $\partial\mathcal{E} \cup \partial\mathcal{S}$.

Discussion. A proof regarding the existence and uniqueness of the equilibrium measure can be found in [ST97, Theorem VIII.1.4], while the statement $\operatorname{supp}(\mu) = \partial\mathcal{E} \cup \partial\mathcal{S}$ follows from the divergence law. As mentioned above, μ describes the charge distribution of a capacitor with a negative unit charge on $\mathcal{E}$ and a charge $\frac{n}{m}$ on $\mathcal{S}$. The assumption regarding the boundedness of $\mathcal{E}$ and $\mathcal{S}$ is required to prevent the charges from "escaping" to infinity. □

We recall from Problem 2.0.1 that $\mathcal{E}$ and $\mathcal{S}$ are assumed to have nonzero logarithmic capacity, which is required in particular for Theorem 2.1.4 to hold. We now clarify the meaning of this statement by defining the complementary class of sets of capacity zero.

Definition 2.1.5 *A set $S \subset \mathbb{C}$ is said to be of (logarithmic) capacity zero if every unit Borel measure μ on S has infinite energy $I(\mu)$ as defined in* (2.4)*. Sets of capacity zero are also called polar.*

We note that discrete sets $S \subset \mathbb{C}$ in particular have capacity zero since any unit measure μ on S must assign a nonzero mass $\mu(\{x\})$ to at least one of the points $x \in S$, and thus the energy $I(\mu)$ contains a term of the form $-\log|x - x| = \infty$. In fact, discrete sets will be the only sets of capacity zero relevant for our purposes.

Theorem 2.1.6 *In the notation of Theorem* 2.1.4*, the equilibrium potential*

$$U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(x) := -\int \log|x - t|\, d\mu_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(t) \tag{2.5}$$

associated with the equilibrium measure $\mu_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}$ satisfies

$$\begin{aligned} \sup_{q.e.\ t\in\mathcal{E}} U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(t) &= U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(x) \qquad \text{for all } x \in \operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}) \cap \mathcal{E}, \\ \inf_{q.e.\ t\in\mathcal{S}} U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(t) &= U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(x) \qquad \text{for all } x \in \operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}) \cap \mathcal{S}, \end{aligned} \tag{2.6}$$

where $\inf_{q.e.\ t\in S} f(t)$ (the quasi-everywhere infimum of $f : S \to \mathbb{R}$ on $S \subset \mathbb{C}$) is defined as the largest constant $L \in \mathbb{R}$ such that the set $\{x \in S \mid f(x) < L\}$ has capacity zero, and likewise for $\sup_{q.e.\ t\in S} f(t)$. Conversely, if the potential $U(x)$ of

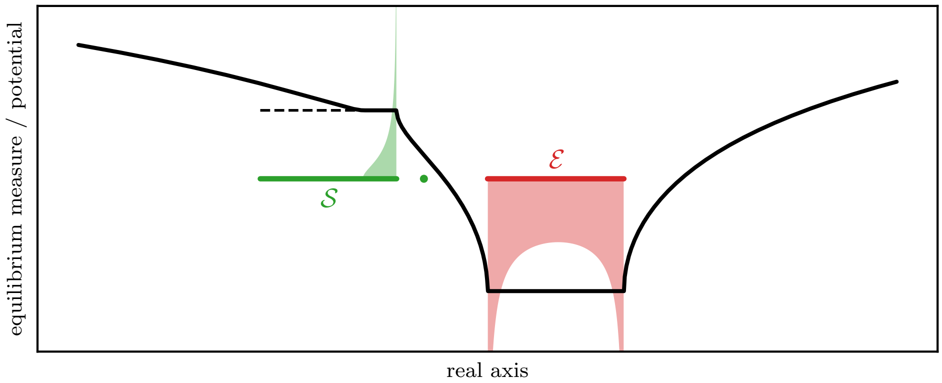


Figure 2.1: Equilibrium measure (shaded) and equilibrium potential (black line) of Example 2.1.7.

a measure μ as described in Theorem 2.1.4 *satisfies* (2.6)*, then μ is the equilibrium measure $\mu_{\mathcal{E},\mathcal{S},n/m}$.*

Discussion. A proof of this statement is given in [ST97, Theorem VIII.2.2]. In physical terms, (2.6) expresses the observation that a tentative charge distribution μ is the equilibrium distribution if and only if there is no way to move charges to an energetically more favourable location. For our purposes, the "quasi everywhere" condition is required to cover the cases where $\mathcal{E}$ or $\mathcal{S}$ contain a set of isolated points S since such a set is too small to hold any charge and hence the potential on S may deviate from the potential on $\operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},n/m})$. □

In most applications, the conditions (2.6) amount to requiring that $U_{\mathcal{E},\mathcal{S},n/m}(x)$ is constant on both $\mathcal{E}$ and $\mathcal{S}$, but there are two important exceptions which are illustrated in the following example.

Example 2.1.7 Figure 2.1 displays the equilibrium measure and potential for

$$\mathcal{E} = [0.5, 2], \qquad \mathcal{S} := [-2, -0.5] \cup \{-0.2\}, \qquad \tfrac{n}{m} = 0.1.$$

We note that the equilibrium potential $U_{\mathcal{E},\mathcal{S},n/m}(x)$ is indeed constant on $\mathcal{E}$, while on $\mathcal{S}$ we have

$$U_{\mathcal{E},\mathcal{S},n/m}(x) \begin{cases} > c & \text{for } x \in [-2,-0.5] \setminus \operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},n/m}), \\ = c & \text{for } x \in [-2,-0.5] \cap \operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},n/m}), \\ < c & \text{for } x = -0.2, \end{cases}$$

where $c := \inf_{\text{q.e. } t\in\mathcal{S}} U_{\mathcal{E},\mathcal{S},n/m}(t)$ is indicated by the dashed line in Figure 2.1. We conclude that even though $U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(x)$ is not constant on $\mathcal{S}$, the energy (2.4) cannot be decreased since moving charge from $\operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}} \cap \mathcal{S})$ to the left would increase the energy and moving charge to $\{-0.2\}$ is not permitted since singletons are too small to hold a charge.

We now return to the problem of optimising the bound (2.3).

Theorem 2.1.8 *In the notation of Theorems* 2.1.2 *and* 2.1.4 *and with*

$$V(\mathcal{E},\mathcal{S},{}^{n}\!/_{m}) := -\sup_{q.e.\ t\in\mathcal{E}} U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(t) + \inf_{q.e.\ t\in\mathcal{S}} U_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}}(t) > 0,$$

we have for all $\alpha = \frac{n}{m} \in [0,\infty)$ *that*

$$\lim_{m\to\infty} \inf_{\substack{X\subset\mathcal{E},\\ \#X=m}} \inf_{\substack{Y\subset\mathcal{S},\\ \#Y=\lfloor\alpha m\rfloor}} \Big(\|\ell_X/\ell_Y\|_{\mathcal{E}}\, \|\ell_Y/\ell_X\|_{\mathcal{S}}\Big)^{1/m} = \exp\big(-V(\mathcal{E},\mathcal{S},\alpha)\big). \tag{2.7}$$

Furthermore, this limit is attained for any sequence X_m, Y_n *such that*

$$\mu_{X_m,Y_n}(x) := -\frac{1}{m}\sum_{k=0}^{m}\delta(x-x_k) + \frac{1}{m}\sum_{\ell=1}^{n}\delta(x-y_\ell) \quad\to\quad \mu_{\mathcal{E},\mathcal{S},{}^{n}\!/_{m}} \tag{2.8}$$

in the weak sense.*

Proof. The above statements are shown in [ST97, §VIII.3]. More precisely, equation (2.7) is shown with $\geq$ instead of $=$ in [ST97, Theorem VIII.3.1], and the sharpness follows from the second part of the statement which is discussed in [ST97, §VIII.3].□

Theorem 2.1.8 combined with the bound (2.3) imply that rational interpolation with well-chosen interpolation points X and poles Y converges exponentially at "essentially" the rate $V(\mathcal{E},\mathcal{S},{}^{n}\!/_{m})$, where the precise meaning of "essentially" will be clarified in Theorem 2.1.10 after introducing the appropriate notation in Definition 2.1.9.

Definition 2.1.9 *A sequence* $a : \mathbb{N} \to [0,\infty)$ *is said to decay exponentially with asymptotic rate* α *if for all* $\tilde{\alpha} < \alpha$ *there exists a constant* $C(\tilde{\alpha})$ *such that* $a_k \leq C(\tilde{\alpha})\exp(-\tilde{\alpha}k)$ *for all* $k \in \mathbb{N}$*. Following the* $\mathcal{O}_\varepsilon$ *notation of [Tre17], we write*

$$a_k \leq_\varepsilon C(\alpha)\exp(-\alpha k) \qquad \text{or} \qquad a_k \lesssim_\varepsilon \exp(-\alpha k)$$

for such sequences.

We note that if $C : [0, \alpha) \to (0, \infty)$ is such that $\lim_{\tilde{\alpha}\to\alpha} C(\tilde{\alpha})$ exists and is bounded, then $a_k \leq_\varepsilon C(\alpha)\, \exp(-\alpha k)$ is equivalent to $a_k \leq C(\alpha)\, \exp\bigl(-\alpha k\bigr)$. A typical example of a sequence $a_k \leq_\varepsilon C(\alpha)\, \exp(-\alpha k)$ is $a_k := k\, \exp(-\alpha k)$, in which case $C(\tilde{\alpha}) = \max_k k\, \exp\bigl(-(\alpha - \tilde{\alpha})\, k\bigr)$ and $\lim_{\tilde{\alpha}\to\alpha} C(\tilde{\alpha}) = \infty$.

Theorem 2.1.10 *Let $f(x)$, $\mathcal{E}$ and $\mathcal{S}$ be as in Problem* 2.0.1*, and denote by $r(x)$ the rational interpolant to $f(x)$ at points X and with poles Y distributed asymptotically according to the equilibrium measure $\mu_{\mathcal{E},\mathcal{S},\frac{n}{m}}$ in the sense of* (2.8)*. We then have that*

$$\|f - r\|_{\mathcal{E}} \lesssim_\varepsilon \exp\bigl(-V(\mathcal{E}, \mathcal{S}, {}^{n}\!/\!{}_{m})\, m\bigr). \tag{2.9}$$

Proof. The idea is to combine the bound from (2.3) with the estimate on $\|\ell_X - \ell_Y\|_{\mathcal{E}}\, \|\ell_Y/\ell_X\|_{\mathcal{S}}$ from Theorem 2.1.8. Details can be found e.g. in [Saf10, §5]. □

Remark 2.1.11 Equation (2.9) holds with only $\lesssim_\varepsilon$ rather than the stronger $\lesssim$ because as mentioned in Remark 2.1.3, the above discussion should have been written in terms of Ω rather $\mathbb{C}\backslash\mathcal{S}$ and Theorem 2.1.10 follows after taking the limit $\Omega \to \mathbb{C}\backslash\mathcal{S}$. As Ω approaches $\mathbb{C} \setminus \mathcal{S}$, the (m, n)-dependent factor $\|\ell_X/\ell_Y\|_{\mathcal{E}}\, \|\ell_Y/\ell_X\|_{\partial\Omega}$ in (2.3) decreases while the prefactor $C \sim \|f\|_{\partial\Omega}$ diverges, which is precisely the behaviour expressed by $\lesssim_\varepsilon$.

Let us conclude this section with a brief summary of the key ideas presented here. We have seen in Theorem 2.1.2 how the interpolation points X and poles Y of a rational interpolant $r(x)$ correspond to negative and positive point charges, respectively, and Theorem 2.1.8 asserted that the rate of convergence of rational interpolation is optimised by choosing $X \subset \mathcal{E}$ and $Y \subset \mathbb{C} \setminus \Omega$ such that the charges are at equilibrium. The key challenge to finding good rational interpolants is thus to determine the equilibrium distribution of X and Y, which is the topic of the next section.

2.2 Determining Equilibrium Measures via Log-Maps

A Borel measure μ and its associated potential $U(x)$ (defined analogously to (2.5)) can be conveniently represented in a single object defined as follows.

Definition 2.2.1 *The log-map $L(x)$ of a finite signed Borel measure μ on $\mathbb{C}$ is the function $L : \mathbb{C} \setminus \operatorname{supp}(\mu) \to \mathbb{C}$ given by*

$$L(x) := -\int \log(x - t)\, \mu(t).$$

While the term “log-map” is our own invention, the function that it refers to has appeared previously in exactly the same context in [ET99, SSW01]. Moreover, if $\mu = \mu_{\mathcal{E},\mathcal{S},0}$ is the equilibrium measure for a simply connected set $\mathcal{E} \subset \mathbb{C}$ in the polynomial case $\frac{n}{m} = 0$, then the associated log-map $L(x)$ satisfies $L(x) = -\log \Phi(x) + \mathrm{const}$ where $\Phi(x)$ is the Riemann map from $\mathbb{C} \setminus \mathcal{E}$ onto the unit disk $\{|z| < 0\}$. Further connections between $L(x)$ and functions from the literature have been pointed out in [ET99, §1].

It follows from the properties of the logarithm that $L(x)$ is locally analytic[1] on its domain of definition and that $\operatorname{Re}\big(L(x)\big)$ is the logarithmic potential $U(x)$ associated with μ, and we will next show that the measure μ itself can be derived from the imaginary part of $L(x)$ under some assumptions which will be satisfied by all the measures μ and log-maps $L(x)$ considered in this thesis.

Theorem 2.2.2 *Let μ and $L(x)$ be as in Definition* 2.2.1*, and assume* $\operatorname{supp}(\mu)$ *has no interior. We then have for all bounded Borel sets $\Omega \subset \mathbb{C}$ with piecewise C^1 boundary that*

$$\mu(\Omega) = -\frac{1}{2\pi i} \int_{\partial\Omega} L'(x)\, dx, \tag{2.10}$$

where $L'(x)$ for points $x \in \operatorname{supp}(\mu)$ outside the domain of analyticity of $L(x)$ is unspecified if x is an isolated or endpoint of $\operatorname{supp}(\mu)$ *(these points have measure zero and hence do not affect the integral), and otherwise defined as*

$$L'(x) := \begin{cases} \lim\limits_{\tilde{x} \to x,\, \tilde{x} \in \Omega} L'(\tilde{x}) & \text{if } x \in \Omega, \\ \lim\limits_{\tilde{x} \to x,\, \tilde{x} \notin \Omega} L'(\tilde{x}) & \text{if } x \notin \Omega. \end{cases}$$

Conversely, a function $L(x)$ is the log-map of a finite signed Borel measure μ if $L(x)$ is locally analytic on $\mathbb{C} \setminus \operatorname{supp}(\mu)$, (2.10) *is satisfied, and we have that*

$$\lim_{|x| \to \infty} L(x) + \mu(\mathbb{C}) \log(x) = 0. \tag{2.11}$$

Proof. We obtain using Cauchy’s integral formula and (2.11) that

$$\mu(\Omega) = \int \chi_\Omega(t)\, d\mu(t) = \frac{1}{2\pi i} \int \int_{\partial\Omega} (x - t)^{-1}\, dx\, d\mu(t) = -\frac{1}{2\pi i} \int_{\partial\Omega} L'(x)\, dx$$

which shows the first part of the theorem. To see the second part, let $\tilde{L}(x)$ be the

[1] A function $f : D \to \mathbb{C}$ with $D \subset \mathbb{C}$ is called locally analytic if for every $x \in D$ there exists a neighbourhood Ω such that f is analytic on Ω. Such functions are also known as multivalued.

log-map of the measure $\mu(\Omega)$. Using the first part of the theorem, it then follows

$$\frac{1}{2\pi i}\int_{\partial\Omega}\big(L'(x)-\tilde{L}'(x)\big)\,dx=\mu(\Omega)-\mu(\Omega)=0$$

which according to Morera's theorem [AG18, Thm. 3.8.10] implies that $L'(x)-\tilde{L}'(x)$ is analytic and has an anti-derivative $L(x)-\tilde{L}(x)$ on $\mathbb{C}$. The claim follows after noting that (2.11) combined with Liouville's theorem [AG18, Thm. 3.9.2] asserts that this anti-derivative vanishes. □

The essence of Theorem 2.2.2 is implicit in the discussion in [ET99, §2-4]. The above clarification has been worked out independently by the author.

To see the connection between (2.10) and the imaginary part of $L(x)$, we note that $\int_{\partial\Omega} L'(x)\,dx$ can be written as

$$\int_{\partial\Omega} L'(x)\,dx = L(x_{\text{end}})-L(x_{\text{start}}) \tag{2.12}$$

where x_{start} denotes some arbitrary point on $\partial\Omega$ and $L(x_{\text{end}})$ denotes the point reached after tracing $L(x)$ for one full revolution along $\partial\Omega$. Equation (2.10) then says that this difference must be equal to $-2\pi i\,\mu(\Omega)$, i.e. $L(\partial\Omega)$ must be a line segment in the complex plane which rises by $2\pi\,|\mu(\Omega)|$ if $\mu(\Omega)$ is negative, and which descends by the same amount if $\mu(\Omega)$ is positive. This is further illustrated in the following example.

Example 2.2.3 Consider the weighted Dirac measure $\mu(x)=\alpha\,\delta(x)$ with $\alpha\in\mathbb{R}$ and its associated log-map $L(x)=-\alpha\log(x)$. If we set $\Omega:=\{z\mid |z|\le 1\}$, then $L(\partial\Omega)$ is a straight line from $\pi i\,\alpha$ to $-\pi i\,\alpha$ which is consistent with Theorem 2.2.2.

Theorem 2.2.2 allows us to determine the equilibrium measure $\mu_{\mathcal{E},\mathcal{S},{}^{n}\!/\!{}_{m}}$ by guessing its associated log-map $L_{\mathcal{E},\mathcal{S},{}^{n}\!/\!{}_{m}}(x)$ and then verifying that the measure and potential resulting from $L_{\mathcal{E},\mathcal{S},{}^{n}\!/\!{}_{m}}(x)$ satisfy the conditions of Theorems 2.1.4 and 2.1.6. The following theorem reformulates these conditions in terms of $L_{\mathcal{E},\mathcal{S},{}^{n}\!/\!{}_{m}}(x)$ under some additional assumptions introduced solely to simplify the exposition. The extension to a more general statement will be obvious.

Theorem 2.2.4 *Assume $\mathcal{E}$ and $\mathcal{S}$ are connected and such that* $\operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},{}^{n}\!/\!{}_{m}})=\partial\mathcal{E}\cup\partial\mathcal{S}$. *Then, a function $L_{\mathcal{E},\mathcal{S},{}^{n}\!/\!{}_{m}}(x)$ is the equilibrium log-map if and only if all of the following conditions are satisfied:*

1. *$L_{\mathcal{E},\mathcal{S},{}^{n}\!/\!{}_{m}}(x)$ is locally analytic on $\mathbb{C}\setminus\big(\mathcal{E}\cup\mathcal{S}\big)$.*

2. $L_{\mathcal{E},\mathcal{S},n/m}(x)$ maps $\partial\Omega$ with $\Omega \in \{\mathcal{E}, \mathcal{S}\}$ to the interval

$$L_{\mathcal{E},\mathcal{S},n/m}\big(\partial\Omega\big) = c_\Omega + i\big[s_\Omega, e_\Omega\big],$$

where $c_\Omega + is_\Omega = L_{\mathcal{E},\mathcal{S},n/m}(x_{\text{start}})$ for some starting point $x_{\text{start}} \in \partial\Omega$ and $c_\Omega + ie_\Omega = L_{\mathcal{E},\mathcal{S},n/m}(x_{\text{end}})$ denotes the point reached after one full revolution along $\partial\Omega$ (cf. (2.12)*).*

3. These intervals are traversed in ascending / descending direction and their lengths are 2π and $2\pi\frac{n}{m}$ for $\Omega = \mathcal{E}$ and $\Omega = \mathcal{S}$, respectively, i.e. $e_\mathcal{E} - s_\mathcal{E} = 2\pi$ and $e_\mathcal{S} - s_\mathcal{S} = -2\pi\frac{n}{m}$.

4. $\lim_{|x|\to\infty} L_{\mathcal{E},\mathcal{S},n/m}(x) + \big(\frac{n}{m} - 1\big)\log(x) = 0$.

Proof. Condition 1 ensures that $L_{\mathcal{E},\mathcal{S},n/m}(x)$ has the analyticity properties required by Theorem 2.2.2, and Condition 2 asserts that the logarithmic potential satisfies the optimality conditions from Theorem 2.1.6 which due to the assumption $\operatorname{supp}(\mu_{\mathcal{E},\mathcal{S},n/m}) = \partial\mathcal{E} \cup \partial\mathcal{S}$ simplify to requiring that $U_{\mathcal{E},\mathcal{S},n/m}(x) = \operatorname{Re}\big(L_{\mathcal{E},\mathcal{S},n/m}(x)\big)$ is constant on $\mathcal{E}$ and $\mathcal{S}$. Condition 3 fixes the charges on $\mathcal{E}$ and $\mathcal{S}$ according to Theorem 2.2.2, and finally Condition 4 ensures that (2.11) is satisfied. □

In the polyomial case $\frac{n}{m} = 0$, Theorem 2.2.4 essentially describes the defining properties of the Green's function with pole at infinity, see e.g. [Saf10, p. 184], [ST97, p. 108] or [Ran95, Def. 4.4.1], and see Remark 2.2.5 below regarding Condition 3 which is not usually listed as a defining property. The extension to the rational case has been derived independently by the author.

Remark 2.2.5 Condition 4 of Theorem 2.2.4 implies Condition 3 in the polynomial case $\frac{n}{m} = 0$ since under these circumstances, the integral $\int_{\partial\mathcal{E}} L'_{\mathcal{E},\mathcal{S},0}(x)\,dx$ can be computed by moving the contour $\partial\mathcal{E}$ far enough into the complex plane such that the difference between $L_{\mathcal{E},\mathcal{S},0}(x)$ and $\log(x)$ becomes negligible and hence

$$\begin{aligned}\int_{\partial\mathcal{E}} L'_{\mathcal{E},\mathcal{S},0}(x)\,dx &= \lim_{r\to\infty}\int_{|x|=r} L'_{\mathcal{E},\mathcal{S},0}(x)\,dx \\ &= \lim_{r\to\infty} L_{\mathcal{E},\mathcal{S},0}\Big(r\,e^{\pi i\,(1-0)}\Big) - L_{\mathcal{E},\mathcal{S},0}\Big(r\,e^{\pi i\,(1+0)}\Big) \\ &= \log\Big(e^{\pi i\,(1-0)}\Big) - \log\Big(e^{\pi i\,(1+0)}\Big) \\ &= 2\pi i.\end{aligned}$$

However, both conditions are required for the rational case $\frac{n}{m} > 0$: without Condition 3, the asymptotic behaviour required by Condition 4 could be achieved simply by scaling the polynomial log-map $L_{\mathcal{E},\mathcal{S},0}(x)$, i.e.

$$\tilde{L}_{\mathcal{E},\mathcal{S},n/m}(x) := \big(1 - \tfrac{n}{m}\big)\, L_{\mathcal{E},\mathcal{S},0}(x)$$

satisfies Conditions 1, 2 and 4 but it is clearly not the correct log-map.

2.3 Polynomial Approximation of the Fermi-Dirac Function

This section demonstrates the application of the above theory by determining in Theorem 2.3.12 the rate of convergence of polynomial approximation to the Fermi-Dirac function $f_{\beta,E_F}(E)$ on the sets

$$\mathcal{E} = [-1,1] \qquad \text{and} \qquad \mathcal{E} = [-1,\varepsilon_-] \cup [\varepsilon_+,1]$$

with $-1 < \varepsilon_- < \varepsilon_+ < 1$, which serve as prototypical examples for the spectra of, respectively, metals and insulators, see Section 1.1. The main step towards this goal is to determine the log-maps for the above sets and ratio $\frac{n}{m} = 0$ between the denominator degree n and numerator degree m, which we will do following the Schwarz-Christoffel mapping techniques of [ET99, SSW01] in Theorems 2.3.4 and 2.3.6.

Before we begin, we would like to point out that for $\frac{n}{m} = 0$, the set of singularities $\mathcal{S}$ disappears from the definition of the equilibrium measure $\mu_{\mathcal{E},\mathcal{S},0}$ in Theorem 2.1.4 and hence we drop the subscripts $\mathcal{S}$ and 0 in $\mu_{\mathcal{E},\mathcal{S},0}$, $U_{\mathcal{E},\mathcal{S},0}(x)$ and $L_{\mathcal{E},\mathcal{S},0}(x)$ in this case. It will further be convenient to have copies $G_{\mathcal{E}}(x)$ and $g_{\mathcal{E}}(x) := \operatorname{Re}\big(G_{\mathcal{E}}(x)\big)$ of, respectively, $L_{\mathcal{E}}(x)$ and $U_{\mathcal{E}}(x)$ at hand which are shifted such that $g_{\mathcal{E}}(x)$ vanishes on $\mathcal{E}$. These are introduced in the following definition.

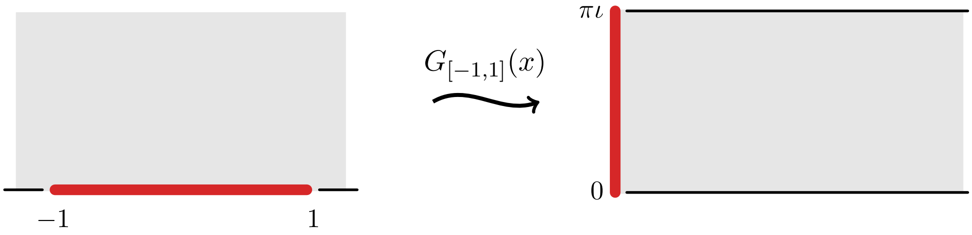
Definition 2.3.1 *We introduce*

$$G_{\mathcal{E}}(x) := L_{\mathcal{E}}(x) - \sup_{q.e.\ t\in\mathcal{E}} U_{\mathcal{E}}(t), \qquad g_{\mathcal{E}}(x) := U_{\mathcal{E}}(x) - \sup_{q.e.\ t\in\mathcal{E}} U_{\mathcal{E}}(t).$$

$g_{\mathcal{E}}(x)$ is known as the Green's function of $\mathcal{E}$ in the literature [Saf10, ST97, Ran95].

Given the above setup, the following observations were made in [ET99, SSW01].

Lemma 2.3.2 *$G_{[-1,1]}(x)$ maps the upper half-plane $\{z \mid \operatorname{Im}(z) > 0\}$ holomorphically onto the semi-infinite strip $\{z \mid \operatorname{Re}(z) > 0, \operatorname{Im}(z) \in [0,\pi)\}$, and we have that $G_{[-1,1]}(-1) = \pi i$, $G_{[-1,1]}(1) = 0$. This map is illustrated in Figure* 2.2.



(a)

(b)

Figure 2.2: Shifted log-map $G_{[-1,1]}(x)$ applied to the upper half-plane (see Lemma 2.3.2 and Theorem 2.3.4).

Proof. Condition 1 of Theorem 2.2.4 asserts that $G_{[-1,1]}(x)$ is analytic on $\mathbb{C} \setminus [-1,1]$, and it follows from Conditions 2 and 3 of the same theorem that $G_{[-1,1]}(x)$ maps $\partial[-1,1]$[2] to $[0, 2\pi i]$. The symmetry of the problem with respect to reflection about the real line further implies that the two segments $G_{[-1,1]}\big([-1,1] \pm 0i\big)$ must be of equal lengths and therefore

$$G_{[-1,1]}\big([-1,1] + 0i\big) = [0, \pi i], \qquad G_{[-1,1]}\big([-1,1] - 0i\big) = [\pi i, 2\pi i]$$

and $G_{[-1,1]}(1) = 0$, $G_{[-1,1]}(-1) = \pi i$, as stated. Another consequence of the aforementioned symmetry is that $U_{[-1,1]}(x) = U_{[-1,1]}(\bar{x})$ and hence

$$0 = \tfrac{\partial}{\partial \operatorname{Im}(x)} U_{[-1,1]}(x) = \tfrac{\partial}{\partial \operatorname{Im}(x)} \operatorname{Re}\big(G_{[-1,1]}(x)\big) = -\tfrac{\partial}{\partial \operatorname{Re}(x)} \operatorname{Im}\big(G_{[-1,1]}(x)\big)$$

for all $x \in \mathbb{R} \setminus [-1,1]$, where the last equality follows from the Cauchy-Riemann equations. This proves that $G_{[-1,1]}\big((-\infty,-1)\big)$ and $G_{[-1,1]}\big((1,\infty)\big)$ must be horizontal lines (the black lines in Figure 2.2b), and it only remains to show that $g_{[-1,1]}(x) = \operatorname{Re}\big(G_{[-1,1]}(x)\big) \to +\infty$ monotonically for $x \to \pm\infty$. The limit follows from Condition 4 of Theorem 2.2.4, and the monotonicity follows after observing that the harmonic and conjugate symmetric function $g_{[-1,1]}(x) = g_{[-1,1]}(\bar{x})$ cannot have a local minimum on $\mathbb{R} \setminus [-1,1]$. □

Figure 2.2 shows that $G_{[-1,1]}(x)$ maps the upper half-plane to a rectangular region where the right edge of the rectangle has been moved to infinity. Functions mapping the upper half-plane holomorphically to a polygon are known as Schwarz-Christoffel maps and have a special structure described in the following theorem.

[2] Recall our convention regarding the boundary contour of curves from Definition 2.0.2.

Theorem 2.3.3 *Let $x_1, \dots, x_n \subset \mathbb{R}$ and let $F(x)$ be a conformal map from the upper half-plane $\{z \mid \operatorname{Im}(z) > 0\}$ to the interior of a polygon P with vertices $F(x_1), \dots, F(x_n)$ and $F(\infty)$. Assume the interior angles at these points are given by $\alpha_1\pi, \dots, \alpha_n\pi$, respectively (the angle at $F(\infty)$ must be $\alpha_\infty := \pi\,\big(n-2-\sum_{k=1}^n \alpha_k\big)$ to ensure that the polygon is closed). We then have*

$$F'(x) = c\prod_{k=1}^{n}(x-x_k)^{\alpha_k-1} \tag{2.13}$$

for some constant $c \in \mathbb{C}$, where here and throughout this chapter, x^α is defined as

$$x^\alpha := |x|^\alpha \exp\big(i\,\alpha\,\arg(x)\big) \qquad \textit{with} \qquad \arg(x) \in (-\pi, \pi]. \tag{2.14}$$

Discussion. A proof of the above result and a detailed discussion of the theory of Schwarz-Christoffel mappings can be found in [DT02]. Briefly, the idea behind (2.13) is that the sign of $F'(x)$ must jump by a factor of $\exp\big(\pi i\,(\alpha_k-1)\big)$ at x_k to generate a vertex of the correct angle, and this is achieved if $F'(x)$ is of the form $F'(x) = f(x)\,(x-x_k)^{\alpha_k-1}$ for some function $f(x)$ which locally has a constant sign.□

In the case of the shifted log-map $G_{[-1,1]}(x)$, the prevertices are $x_1 = -1$, $x_2 = 1$ and the associated angles are $\alpha_1 = \alpha_2 = \frac{1}{2}$; hence we have that

$$G'_{[-1,1]}(x) = \frac{c}{\sqrt{x+1}\,\sqrt{x-1}}.$$

An anti-derivative of this function is given by

$$G_{[-1,1]}(x) = \log\big(x + \sqrt{x+1}\sqrt{x-1}\big), \tag{2.15}$$

where the additive constant has been chosen such that $G_{[-1,1]}(1) = 0$ and the multiplicative constant c has been fixed to satisfy $G_{[-1,1]}(-1) = \pi i$ as required by Lemma 2.3.2. This proves the following well-known result, see e.g. [Saf10, Example 1.11].

Theorem 2.3.4 *The shifted log-map for the interval $[-1,1]$ is given by $G_{[-1,1]}(x) = \log\big(x + \sqrt{x+1}\sqrt{x-1}\big)$.*

Remark 2.3.5 In expressions like (2.15), it is tempting to replace $\sqrt{x+1}\,\sqrt{x-1}$ with $\sqrt{x^2-1}$, but this is not correct for $\operatorname{Re}(x) < 0$ since there the two expressions evaluate different branches of the same function, cf. (2.14).

Let us now extend the above construction to the case of two intervals $\mathcal{E} =$

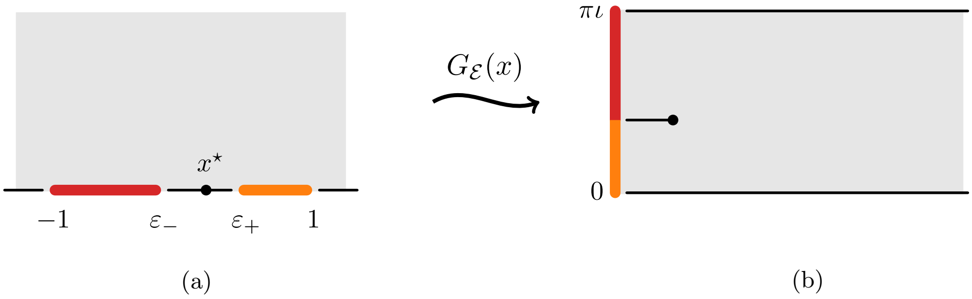


Figure 2.3: Shifted log-map $G_{\mathcal{E}}(x)$ with $\mathcal{E} := [-1, \varepsilon_-] \cup [\varepsilon_+, 1]$ applied to the upper half-plane (see Theorem 2.3.6).

$[-1, \varepsilon_-] \cup [\varepsilon_+, 1]$. Arguing similarly as in Lemma 2.3.2, one can show that the image of $G_{\mathcal{E}}(x)$ applied to the upper half-plane must be of the form shown in Figure 2.3b, where the main novelty is that the polygon on the right-hand side has one vertex (indicated by the black dot) whose preimage x_k is not an endpoint of $\mathcal{E}$ but rather some point $x^\star \in (\varepsilon_-, \varepsilon_+)$, and whose interior angle α_k is 2 rather than $\frac{1}{2}$. This additional vertex is required to ensure that the two segments of $\operatorname{Im}\bigl(G_{\mathcal{E}}(\mathcal{E})\bigr)$ (the red and orange lines in Figure 2.3b) fall on a single vertical line, and the preimage $x^\star$ is determined so as to satisfy this condition. Using the Schwarz-Christoffel formula from Theorem 2.3.3, this yields the following shifted log-map.

Theorem 2.3.6 ([SSW01]) *The shifted log-map for* $\mathcal{E} := [-1, \varepsilon_-] \cup [\varepsilon_+, 1]$ *with* $-1 < \varepsilon_- < \varepsilon_+ < 1$ *is given by*

$$G_{\mathcal{E}}(x) = \int_1^x f(t)\,(t - x^\star)\,dt \tag{2.16}$$

where

$$f(x) := \frac{1}{\sqrt{x+1}\,\sqrt{x-\varepsilon_-}\,\sqrt{x-\varepsilon_+}\,\sqrt{x-1}} \qquad \textit{and} \qquad x^\star := \frac{\int_{\varepsilon_-}^{\varepsilon_+} t\,f(t)\,dt}{\int_{\varepsilon_-}^{\varepsilon_+} f(t)\,dt}.$$

Proof. The functional form of $G_{\mathcal{E}}(x)$ follows immediately from Figure 2.3 and Theorem 2.3.3 (Schwarz-Christoffel mapping), and we observe that

$$\lim_{|x|\to\infty} G_{\mathcal{E}}(x) - \log(x) = \lim_{|x|\to\infty} \int_1^x \frac{t-x^\star}{\sqrt{t+1}\,\sqrt{t-\varepsilon_-}\,\sqrt{t-\varepsilon_+}\,\sqrt{t-1}} - \tfrac{1}{t}\,dt = \text{const};$$

hence $G_{\mathcal{E}}(x)$ satisfies Condition 4 of Theorem 2.2.4 which in turn implies Condition 3 of the same theorem, see Remark 2.2.5. It remains to show that $G_{\mathcal{E}}(\mathcal{E})$ falls on a

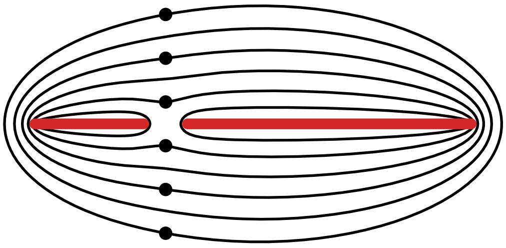


Figure 2.4: Equipotential lines $\{x \mid g_{\mathcal{E}}(x) = \text{const}\}$ of the Green's function for $\mathcal{E} = [-1, -0.5] \cup [-0.3, 1]$. The black dots indicate the singularities of the Fermi-Dirac function for $E_F = 0.4$ and $\beta = \frac{\pi}{10}$.

single vertical line, which we conclude from Figure 2.3b to be equivalent to

$$\int_{\varepsilon_-}^{x^\star} f(t)\,(t - x^\star)\,dt = -\int_{x^\star}^{\varepsilon_+} f(t)\,(t - x^\star)\,dt.$$

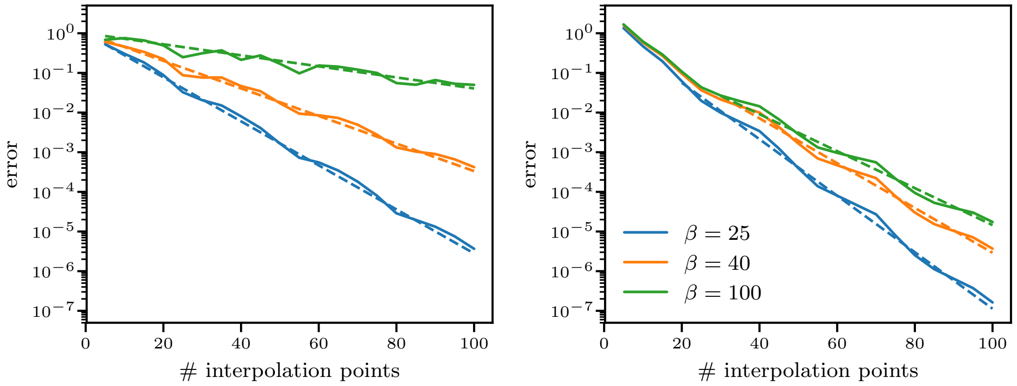
Upon rearranging, this identity becomes the defining formula for $x^\star$ and hence this condition is indeed satisfied. □

Remark 2.3.7 The standard software package for evaluating integrals of the form (2.16) is the Schwarz-Christoffel toolbox from [Dri96]. However, the numerical experiments reported below are based on our own code which is available online at `github.com/ettersi/SchwarzChristoffel.jl` and which employs techniques described in [DT02].

We recall that our motivation for determining log-maps was Theorem 2.1.10, which lower-bounds the rate of convergence of rational interpolation by the difference in potential $V(\mathcal{E}, \mathcal{S}, n/m)$ between $\mathcal{E}$ and $\mathcal{S}$. In the polynomial case $\frac{n}{m} = 0$, this difference in potential is given by

$$V(\mathcal{E}, \mathcal{S}, 0) = \inf_{x \in \mathcal{S}} g_{\mathcal{E}}(x),$$

which is the value V determining the largest level set $\big\{x \mid g_{\mathcal{E}}(x) < V\big\}$ contained in the domain of analyticity of $f(x)$, see Figure 2.4. Furthermore, this lower bound $V(\mathcal{E}, \mathcal{S}, 0)$ is in fact the exact rate of convergence, and it is the exact rate of convergence not only for interpolation but also for best approximation. This observation is a consequence of the following result, which states that if a sequence of polynomials



(a) Metal

(b) Insulator

Figure 2.5: Convergence of polynomial interpolation to the Fermi-Dirac function $f_{\beta,E_F}(E)$ with $E_F = -0.2$ and $\mathcal{E} = [-1,1]$ (metal, left) and $\mathcal{E} = [-1,-0.3] \cup [-0.1,1]$ (insulator, right). The dashed lines indicate the rate of convergence $g_{\mathcal{E}}\big(E_F + \frac{\pi i}{\beta}\big)$ predicted by Theorem 2.3.9.

$p_n(x) \in \mathcal{P}_n$ converges to a function $f(x)$ with a rate faster than $V(\mathcal{E}, \mathcal{S}, 0)$, then $f(x)$ must be analytic on a domain strictly larger than $\{x \mid g_{\mathcal{E}}(x) < V(\mathcal{E}, \mathcal{S}, 0)\}$.

Theorem 2.3.8 *Let $f(x)$ and $\mathcal{E}$ be as in Problem 2.0.1 and $V > 0$. The following statements are equivalent.*

- $\inf\limits_{p \in \mathcal{P}_m} \|f - p\|_{\mathcal{E}} \lesssim_{\varepsilon} \exp\big(-V\, m\big)$
- *$f(x)$ is analytic on $\big\{x \mid g_{\mathcal{E}}(x) < V\big\}$.*

Proof. See e.g. [Saf10, Theorem 4.1] or [Tre13, §8]. □

In the context of electronic structure algorithms, we are interested in the rate of convergence of polynomial approximation to the Fermi-Dirac function $f_{\beta,E_F}(E)$ from (1.4). Theorem 2.3.8 combined with the set of singularities $\mathcal{S}\big(f_{\beta,E_F}\big)$ from (2.1) yields the following result.

Theorem 2.3.9 *The optimal rate of convergence of polynomial approximation to the Fermi-Dirac function, i.e. the largest V such that*

$$\inf_{p \in \mathcal{P}_m} \|f - p\|_{\mathcal{E}} \lesssim_{\varepsilon} \exp\big(-V\, m\big),$$

is given by

$$V(\mathcal{E}, \mathcal{S}_{\beta,E_F}, 0) = g_{\mathcal{E}}\big(E_F + \tfrac{\pi i}{\beta}\big).$$

In order to achieve the optimal rate of convergence from Theorem 2.3.9 through interpolation, we must devise a procedure to compute interpolation points X distributed according to the equilibrium measure $\mu_{\mathcal{E}}$, cf. Theorem 2.1.10. Such a procedure is described in the following theorem.

Theorem 2.3.10 *The points*

$$X := G_{\mathcal{E}}^{-1}\Big(\tfrac{2\pi i}{n-1}\,\{0,\dots,n-1\}\Big)$$

are asymptotically distributed according to the equilibrium measure $\mu_{\mathcal{E}}$ in the sense of (2.8)*.*

Proof. It follows from Theorem 2.2.2 that

$$-\frac{1}{2\pi}\,\Big(\mathrm{Im}\big(L_{\mathcal{E}}(E)\big) - s_{\mathcal{E}}\Big)$$

is the cumulative distribution function (CDF) of $\mu_{\mathcal{E}}$ on $\partial\mathcal{E}$. The above construction is thus equivalent to the technique of simulating a random variable by inverting its CDF, see e.g. [Dev06, §1.1]. □

Example 2.3.11 One may verify by straightforward computation that the inverse to the log-map $G_{[-1,1]}(x)$ from Theorem 2.3.4 is given by $G_{[-1,1]}^{-1}(z) = \cosh(z)$; hence the points X constructed according to Theorem 2.3.10 are given by

$$X = \cosh\Big(\tfrac{2\pi i}{n-1}\,\{0,\dots,n-1\}\Big) = \cos\Big(\tfrac{2\pi}{n-1}\,\{0,\dots,n-1\}\Big).$$

We observe that for odd n, each point in X appears twice, which is a consequence of $\mu_{\mathcal{E}}$ being supported on $\partial\mathcal{E}$ even if $\mathcal{E}$ is a curve (recall our convention regarding the boundary of curves from Definition 2.0.2). Strictly speaking, half of the points in the set X given above are thus located on $[-1,1]+0i$ while the other half live on $[-1,1]-0i$, but of course this technical subtlety does not make any difference in actual computations. The double-sampling on sets $\mathcal{E}$ which are curves is often undesirable since the resulting points X are less evenly distributed on $\mathcal{E}$ than they could be, and the interpolation problem becomes ill-defined if two points $x_1, x_2 \in X$ coincide. These issues may be avoided by sampling according to the averaged equilibrium measure

$$d\bar{\mu}_{\mathcal{E}}(x) := \frac{d\mu_{\mathcal{E}}(x+0i) + d\mu_{\mathcal{E}}(x-0i)}{2},$$

which was done in all the numerical experiments reported in this chapter. If we use this trick to determine interpolation points for the interval $\mathcal{E} = [-1, 1]$, we obtain the well-known Chebyshev points

$$X = \cos\bigl(\tfrac{\pi}{n-1}\{0, \dots, n-1\}\bigr),$$

see e.g. [Tre13, eq. (2.1)].

Figure 2.5 compares the rates of convergence for interpolation in points X generated as described above against the predictions of Theorem 2.3.9, and we observe that theory and experiment match perfectly.

We conclude this section by establishing the asymptotic behaviour of the convergence rate $g_{\mathcal{E}}\bigl(E_F + \frac{\pi i}{\beta}\bigr)$ in the limits of vanishing temperature β^{-1}, vanishing band gap $\delta E = \varepsilon_+ - \varepsilon_-$ (equations (2.17) and (2.18), respectively), and p-doped and n-doped semiconductors (equations (2.19) and (2.20), respectively) where the Fermi-level approaches one of the ends of the band gap. These formulae are useful e.g. for estimating the costs of electronic structure algorithms based on polynomial approximation and are also connected with the localisation of the density matrix discussed in Section 1.3 as we shall see in the next chapter. Formula (2.17) has appeared previously in [BBR13], while the other formulae are new to the best of the author's knowledge.

Theorem 2.3.12 *We have for* $-1 < \varepsilon_- < E_F < \varepsilon_+ < 1$ *that*

$$\begin{aligned}
&g_{[-1,1]}\bigl(E_F + \tfrac{\pi i}{\beta}\bigr) \sim \beta^{-1} && \text{for } \beta \to \infty, && (2.17)\\
&g_{[-1,\varepsilon_-]\cup[\varepsilon_+,1]}\bigl(E_F\bigr) \sim \varepsilon_+ - \varepsilon_- && \text{for } \varepsilon_-, \varepsilon_+ \to E_F, && (2.18)\\
&g_{[-1,\varepsilon_-]\cup[\varepsilon_+,1]}\bigl(E_F\bigr) \sim \sqrt{E_F - \varepsilon_-} && \text{for } E_F \to \varepsilon_-, && (2.19)\\
&g_{[-1,\varepsilon_-]\cup[\varepsilon_+,1]}\bigl(E_F\bigr) \sim \sqrt{\varepsilon_+ - E_F} && \text{for } E_F \to \varepsilon_+. && (2.20)
\end{aligned}$$

In (2.18)*, it is assumed that the limit is approached symmetrically, i.e.* $E_F - \varepsilon_- \sim \varepsilon_+ - E_F$*. The notation* $f(x) \sim g(x)$ *is defined in Appendix* A.2.

Proof. Equations (2.17), (2.19), and (2.20) follow immediately from Theorems 2.3.4 and 2.3.6 by integrating the known derivatives of $G_{\mathcal{E}}(x)$ starting from the limit points. Equation (2.18) follows after noting that the integral

$$\int_{\varepsilon_-}^{E_F} \frac{dt}{\sqrt{t+1}\sqrt{t-\varepsilon_-}\sqrt{t-\varepsilon_+}\sqrt{t-1}}$$

converges to a finite and nonzero limit for $\varepsilon_-, \varepsilon_+ \to E_F$ due to the inverse-square-root singularities at the points $t \in \{\varepsilon_-, \varepsilon_+\}$, and weighting this integral with $t - x^\star$

reduces it to within a constant factor of $\varepsilon_+ - \varepsilon_-$ since $x^\star \in (\varepsilon_-, \varepsilon_+)$. □

2.4 Rational Approximation of the Fermi-Dirac Function

Let us now turn our attention to the rational approximation of the Fermi-Dirac function $f_{\beta,E_F}(E)$ from (1.4). The theory presented in Section 2.1 suggests that we construct such approximations through interpolation, but the application of Theorem 2.1.10 faces the challenge that the set of singularities $\mathcal{S}_{\beta,E_F}$ is discrete; hence it is polar and the equilibrium measure problem associated with $\mathcal{E}$, $\mathcal{S}_{\beta,E_F}$ and $\frac{n}{m} > 0$ is ill-posed. In practical terms, this means that rational approximation to the Fermi-Dirac function can achieve arbitrarily large convergence rates as we will demonstrate next.

Theorem 2.4.1 *Let $m, n \in \mathbb{N}$ with n even, and let $\mathcal{E} \subset \mathbb{R}$ be a non-polar set. We introduce*

$$q_n(E) := -\frac{1}{\beta} \sum_{y \in Y_n} \frac{1}{E - y}$$

with $Y_n := \big\{ E_F + \frac{\pi i k}{\beta} \;\big|\; k \in \{-2n+1, -2n-1, \dots, 2n-1\} \big\} \subset \mathcal{S}_{\beta,E_F}$ and set

$$r_{mn}(E) := p_m(E) + q_n(E)$$

where $p_m \in \mathcal{P}_m$ denotes a polynomial interpolant to $f_{\beta,E_F}(E) - q_n(E)$ with interpolation points distributed according the equilibrium measure $\mu_{\mathcal{E}}$. We then have that

$$\|r_{mn} - f\|_{\mathcal{E}} \lesssim_{\mathcal{E}} \exp\Big(-g_{\mathcal{E}}\Big(E_F + \tfrac{\pi i\,(2n+1)}{\beta}\Big)\, m \Big);$$

hence the convergence rate for $m \to \infty$ can be made arbitrarily large by choosing n large enough.

Proof. For each $z \in \mathcal{S}_{\beta,E_F}$, we compute using L'Hôpital's rule that

$$\lim_{x\to y} (x-y)\, f_{\beta,E_F}(x) = \lim_{x \to y} \tfrac{x-y}{1+\exp\big(\beta\,(E-E_F)\big)} = \lim_{x\to y} \tfrac{1}{\beta\, \exp\big(\beta\,(E-E_F)\big)} = -\frac{1}{\beta};$$

hence we conclude that the singularities in S_{β,E_F} are simple poles with residues $-\frac{1}{\beta}$. The poles of the two functions $f_{\beta,E_F}(E)$ and $q_n(E)$ at $E \in Y_n$ thus cancel in the polynomial approximation problem $p_m(E) \approx f_{\beta,E_F}(E) - q_n(E)$ such that the claim follows from the polynomial version of Theorem 2.1.10.

The idea of subtracting poles from the Fermi-Dirac function to increase its domain of analyticity originated in the master thesis of Matthew Coates [Coa18] who was supervised by the author of the present thesis. □

Remark 2.4.2 The above discussion might seem to suggest the implication

$$\mathcal{S} \text{ is polar} \implies \text{rational approximation converges superexponentially},$$

but this is correct if and only if all the singularities in $\mathcal{S}$ are poles. For essential singularities, the Laurent series has infinitely many negative powers and hence we cannot subtract poles as in Theorem 2.4.1.

Theorem 2.4.1 makes a strong theoretical point by showing that rational approximation to the Fermi-Dirac function may converge superexponentially, but its practical relevance is limited for the following reason. We have seen in Theorem 2.3.12 that $g_{[-1,1]}(x) \sim |\operatorname{Im}(x)|$ for x approaching $(-1,1)$; hence in the regime $n \ll \beta$ we have the estimate

$$\|r_{mn} - f\|_{[-1,1]} \lesssim_\varepsilon \exp\bigl(-C\,\tfrac{mn}{\beta}\bigr)$$

for some $C > 0$. This shows that at least one of m or n must scale algebraically in β in order to achieve a constant error, which is much worse than it could be as we shall see shortly.

The reason for the poor scaling of the approximation scheme from Theorem 2.4.1 is that in the limit $\beta \to \infty$, the Fermi-Dirac function $f_{\beta,E_F}(E)$ degenerates into a step function

$$f_{\infty,E_F}(E) = \begin{cases} 1 & \text{if } \operatorname{Re}(E) < E_F, \\ 0 & \text{if } \operatorname{Re}(E) > E_F \end{cases}$$

which has no isolated singularities and hence the pole-removal trick from Theorem 2.4.1 no longer works. In this regime, more effective rational approximations can be obtained by replacing the exact set of singularities $\mathcal{S}_{\beta,E_F}$ with the set

$$\mathcal{S}_{\beta,E_F}^{\text{filled}} = E_F + \bigl(-\infty i, -\tfrac{\pi i}{\beta}\bigr] \cup \bigl[\tfrac{\pi i}{\beta}, \infty i\bigr)$$

where all gaps between the poles of $\mathcal{S}_{\beta,E_F}$ have been filled in except around the real axis. This set $\mathcal{S}_{\beta,E_F}^{\text{filled}}$ is no longer polar and hence we can apply the rational interpolation theory from Section 2.1 once we have determined the log-map

$$G_{\mathcal{E},\beta,E_F}(x) := G_{\mathcal{E},\mathcal{S}_{\beta,E_F}^{\text{filled}},1}(x).$$

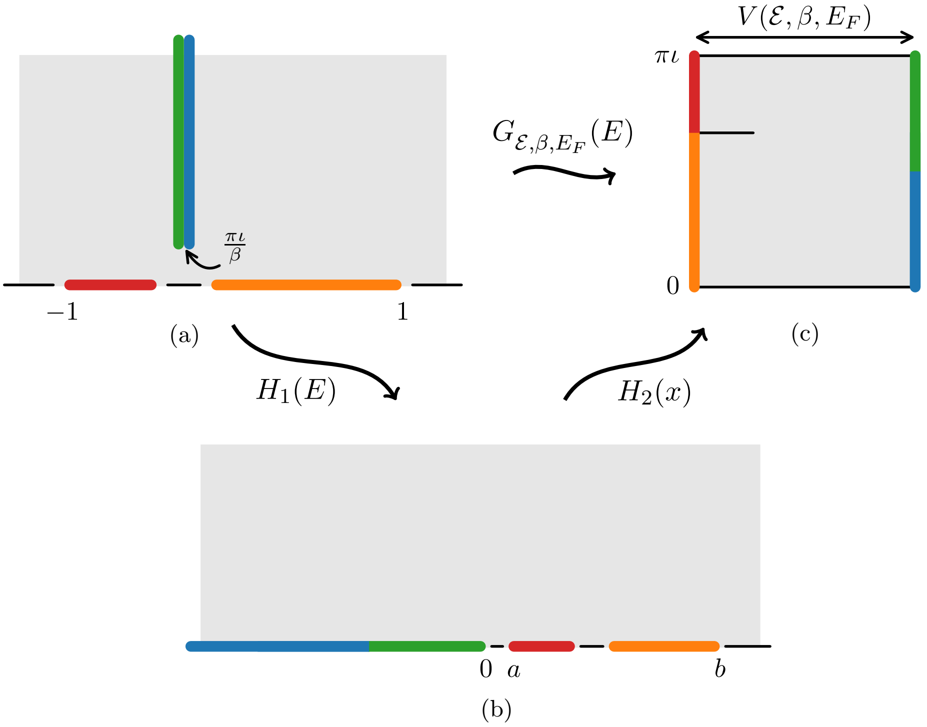


Figure 2.6: Shifted log-map $G_{\mathcal{E},\beta,E_F}(E)$ with $\mathcal{E} = [-1,\varepsilon_-] \cup [\varepsilon_+,1]$ applied to the upper half-plane (see Theorem 2.4.3).

As indicated, we will only consider the case $\frac{n}{m} = 1$ since this is the ratio relevant for the pole-expanded ansatz (1.16) required by the PEXSI algorithm, and for simplicity we only discuss the case $\mathcal{E} = [-1,1]$ in Theorem 2.4.3 below. The modifications required for insulator spectra $\mathcal{E} = [-1,\varepsilon_-] \cup [\varepsilon_+,1]$ are analogous to Theorem 2.3.6 and are illustrated in Figure 2.6.

Theorem 2.4.3 *Let $\beta \in (0,\infty)$ and $E_F \in (-1,1)$. We then have that*

$$G_{[-1,1],\beta,E_F}(E) = \alpha\, H_2\Big(H_1\big(E - E_F\big)\Big) \qquad \text{with} \qquad \alpha := \frac{\pi}{\big|H_2\big(H_1(-1-E_F)\big)\big|},$$

where

$$H_1(E) := \frac{E^2 + E\sqrt{E^2+4c}}{2} + c = \frac{-2cE}{E - \sqrt{E^2+4c}} + c \tag{2.21}$$

with $c = \frac{\pi^2}{4\beta^2}$ is the inverse of $h(x) := \frac{x-c}{\sqrt{x}}$, and

$$H_2(x) := \int_b^x \frac{1}{\sqrt{t}\,\sqrt{t-a}\,\sqrt{t-b}}\,dt \tag{2.22}$$

with $a = H_1(-1)$, $b = H_1(1)$. We provide two formulae for $H_1(E)$ since the first formula in (2.21) *suffers from cancellation in the limit $E \to -\infty$, while the second is numerically unstable in the limit $E \to +\infty$.*

Proof. As in Theorems 2.3.4 and 2.3.6, the proof amounts to showing that the given function $G_{[-1,1],\beta,E_F}(E)$ maps the upper half-plane to the appropriate rectangular region, which in this case is as shown in Figure 2.6c. We note that this rectangle is bounded since according to Condition 4 of Theorem 2.2.4, we must have for $\frac{n}{m} = 1$ that

$$\lim_{|x|\to\infty} L_{\mathcal{E},\mathcal{S},1}(x) = 0.$$

One may verify that $h(x)$ is a holomorphic map from the upper half-plane $H^+ := \{x \mid \operatorname{Im}(x) > 0\}$ to the slit half-plane $H^+ \setminus \big[\frac{\pi i}{\beta}, \infty i\big)$ and maps the boundary $\mathbb{R}$ according to

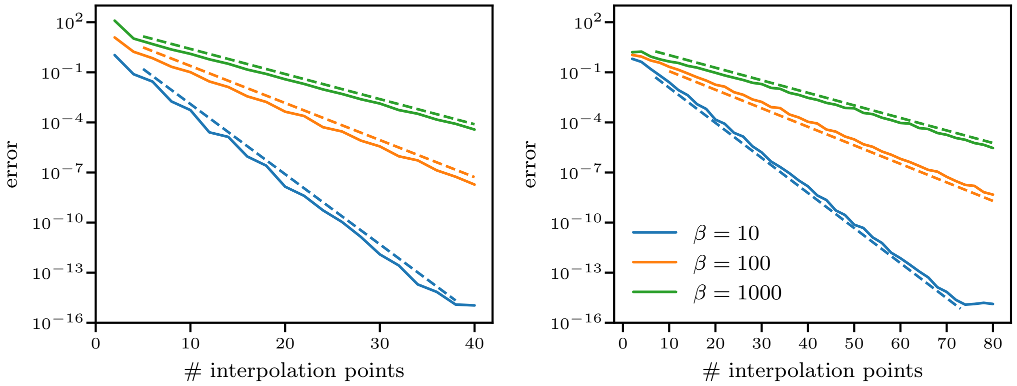
$$h\big((0,\infty)\big) = \mathbb{R}, \qquad h\big([-c,0)\big) = \big[\tfrac{\pi i}{\beta}, \infty i\big) - 0, \qquad h\big((-\infty,-c]\big) = \big[\tfrac{\pi i}{\beta}, \infty i\big) + 0;$$

hence its inverse $H_1(E)$ is as shown in Figures 2.6a and 2.6b. This map was chosen such that the mapping between Figures 2.6b and 2.6c is of Schwarz-Christoffel form as described in Theorem 2.3.3, and the $H_2(x)$ given above immediately follows from this observation. Finally, we remark that the scaling α is needed to ensure $G_{\mathcal{E},\beta,E_F}(-1) = \pi i$ as required by Theorem 2.2.4. □

Please see Remark 2.3.7 for some comments regarding the practical evaluation of the shifted log-map $G_{\mathcal{E},\beta,E_F}(E)$ from Theorem 2.4.3.

Remark 2.4.4 A similar mapping from the slit upper half-plane to a rectangle has been proposed in [LLYE09]. Like our construction, the map given there first maps the boundary segments onto the real line and then uses a second map to "fold" the real line into a rectangle. This second map from [LLYE09] has been derived in [HHT08], and while the construction there may superficially look very different from our formula for $H_2(x)$, the two functions are in fact equivalent since both map the same intervals on the real line onto equivalent edges of the rectangle.

The first map from [LLYE09] is given by $\tilde{H}_1(E) := E^2 + \frac{\pi^2}{\beta^2}$, and this function is genuinely different from our choice in Theorem 2.4.3. We prefer our construction since $\tilde{H}_1(E)$ maps the two points $E \in \mathbb{R}$ and $-E$ to the same image which effectively forces $\mathcal{E}$ to be symmetric with respect to the Fermi level E_F. However, if this symmetry is already satisfied then our map $G_{\mathcal{E},\beta,E_F}(E)$ becomes once again equivalent to its counterpart from [LLYE09] since it performs an equivalent mapping of the boundary.



(a) Interpolation

(b) Contour quadrature

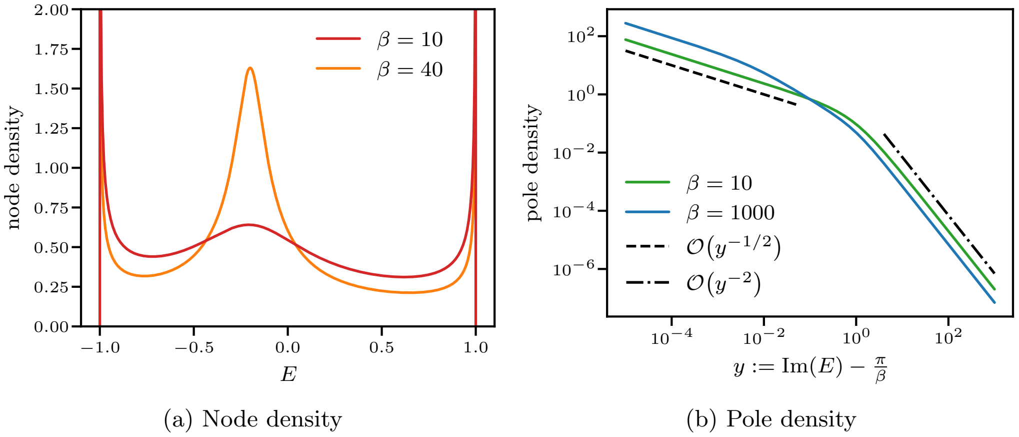
Figure 2.7: Convergence of rational interpolation (left) and contour quadrature (right) for $\mathcal{E} = [-1, 1]$ and $E_F = -0.2$. The dashed lines indicate the rates of convergence $V(\mathcal{E}, \beta, E_F)$ and $\frac{1}{2}\, V(\mathcal{E}, \beta, E_F)$ predicted by Theorem 2.1.10 and (2.28), respectively. Figure (b) will be discussed in Section 2.5.

Theorem 2.4.3 allows us to determine “good” interpolation points X and poles Y in the sense of Theorem 2.1.10 as described in Theorem 2.3.10, and Figure 2.7a demonstrates that the resulting rational interpolants indeed converge at the rate

$$V(\mathcal{E}, \beta, E_F) := V\big(\mathcal{E}, \mathcal{S}^{\text{filled}}_{\beta,E_F}, 1\big) \tag{2.23}$$

predicted by Theorem 2.1.10. For the computations in Figure 2.7a, we represented these interpolants in the pole-expanded form from (1.16) with coefficients determined by inverting the Cauchy matrix associated with the interpolation points and poles, and since this scheme achieves accuracies close to machine precision we conclude that it is robust against rounding errors. Numerical stability of high-order polynomial and rational approximation algorithms is a well-known issue, see e.g. the discussion in [Tre13, §14], and it is not obvious why the scheme described above would not suffer from such instabilities. This will be further investigated in future work.

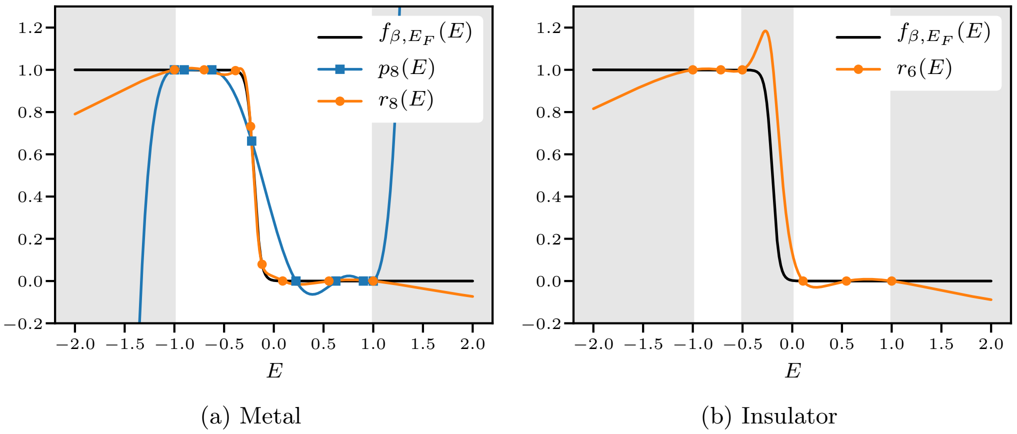
Figure 2.8a shows that for growing β, the interpolation points resulting from Theorem 2.4.3 increasingly concentrate around the Fermi level E_F, and we see in Figure 2.9a that this concentration allows rational interpolants to resolve the jump in the Fermi-Dirac function much better than polynomial interpolants. Figure 2.9b further demonstrates that rational interpolants to $f_{\beta,E_F}(E)$ on gapped spectra $\mathcal{E}$ genuinely approximate only on $\mathcal{E}$ and may significantly deviate from $f_{\beta,E_F}(E)$ on the gap. We conclude from these examples that rational interpolation samples the



(a) Node density

(b) Pole density

Figure 2.8: Density of interpolation points (left) and poles (right) according to the equilibrium measure $\mu_{\mathcal{E},\beta,E_F}$ derived from Theorem 2.4.3 with $\mathcal{E} = [-1, 1]$ and $E_F = -0.2$.



(a) Metal

(b) Insulator

Figure 2.9: Polynomial and rational interpolants of the Fermi-Dirac function $f_{\beta,E_F}(E)$ with $\beta = 20$, $E_F = -0.2$, and $\mathcal{E} = [-1, 1]$ (metal, left) and $\mathcal{E} = [-1, -0.5] \cup [0, 1]$ (insulator, right).

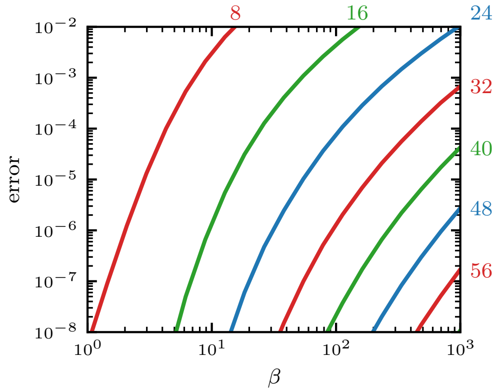


Figure 2.10: Approximation error $\|f_{\beta,0} - r_n\|_{[-1,1]}$ as a function of β for the rational interpolants r_n based on log-map from Theorem 2.4.3. Labels denote the number of interpolation points and poles.

approximand $f(x)$ adaptively in the sense that $f(x)$ is evaluated only in regions where accuracy is required and $f(x)$ lacks smoothness. Finally, Figure 2.8b shows that the pole density prescribed by Theorem 2.4.3 has inverse-square-root singularities at the endpoints $E = E_F \pm \frac{\pi i}{\beta}$ of $\mathcal{S}_{\beta,E_F}^{\text{filled}}$ and decays like $\mathcal{O}\big(E^{-2}\big)$ for $E \to E_F \pm \infty i$.

Figure 2.10 displays the error in the rational interpolants constructed according to the log-map from Theorem 2.4.3 as a function of β and the number of interpolation points and poles n. Comparing this figure against the analogous plots in [Mou16, Fig. 3], we conclude that our rational interpolation scheme performs about a factor of two worse than the best rational approximations constructed there. This difference of a factor two between the optimal convergence rate of rational approximation and the one predicted by logarithmic potential theory has previously been observed in the literature, see the discussion and references at the end of [Saf10, §5], and we plan to further investigate this phenomenon in future work. We remark that Figure 2.10 extends only up to $\beta = 10^3$ while the plots in [Mou16] extend up to $\beta = 10^6$ since we faced numerical challenges when evaluating the integral (2.22) in the limit $\beta \to \infty$ where $a = H_1(-1) \to 0$ but $b = H_1(1) \to 1$. We would like to emphasise that this issue is particular to our implementation and can be overcome either by using more sophisticated Schwarz-Christoffel mapping techniques as discussed in [DT02], or by solving the equilibrium measure problem using the tools from [Olv11].

As in Theorem 2.3.12, we would like to conclude this section by establishing the asymptotic dependence of the convergence rate $V(\mathcal{E}, \beta, E_F)$ from (2.23) on the inverse temperature β and band gap δE. However, unlike in Theorem 2.3.12, we allow only $E_F = 0$ in (2.25), and we do not discuss the case of doped semiconductors since the log-map $G_{\mathcal{E},\beta,E_F}(E)$ from Theorem 2.4.3 does not easily extend to the case

$\beta = \infty$ if $\mathcal{E}$ is not symmetric around the Fermi level E_F (further details will be given in the proof). We believe that this is a purely technical obstacle, and we expect logarithmic dependence on the relevant parameters in all of the aforementioned cases.

Equivalent statements regarding the asymptotic behaviour of the convergence rate of rational approximation to the Fermi-Dirac function have appeared previously in [LLYE09]. The novelty in the theorem below is that it establishes this known result in a new and more general framework.

Theorem 2.4.5 *We have for $-1 < E_F < 1$ that*

$$V\big([-1,1],\beta,E_F\big) \sim \log(\beta)^{-1} \qquad \textit{for } \beta \to \infty, \tag{2.24}$$

$$V\big([-1,-\varepsilon]\cup[\varepsilon,1],\infty,0\big) \sim |\log(\varepsilon)|^{-1} \qquad \textit{for } \varepsilon \to 0. \tag{2.25}$$

The notation $f(x) \sim g(x)$ is defined in Appendix A.2.

Proof. In both cases, the rate of convergence equals the width of the rectangle in Figure 2.6c which is given by

$$V(\mathcal{E},\beta,E_F) = \alpha\,\big|H_2(a) - H_2(0)\big| = \frac{\left|\int_0^a \frac{1}{\sqrt{t}\,\sqrt{t-a}\,\sqrt{t-b}}\,dt\right|}{\left|\int_a^b \frac{1}{\sqrt{t}\,\sqrt{t-a}\,\sqrt{t-b}}\,dt\right|}. \tag{2.26}$$

We analyse this ratio separately for the two cases.

(2.24): For $\beta \to \infty$, the point $a = H_1(-1)$ in (2.21) approaches 0 while $b \to 1$; hence the numerator in (2.26) approaches a finite and nonzero limit due to the inverse-square-root singularities at the endpoints of the integral. The denominator, on the other hand, behaves asymptotically like

$$\left|\int_a^1 \frac{1}{t\,\sqrt{t-1}}\,dt\right| \sim |\log(a)|$$

and thus the claim follows after noting that $a = H_1(-1) \sim \beta^{-4}$, cf. (2.21).

(2.25): For $\beta \to \infty$ or equivalently $c \to 0$, the formula for $G_{\mathcal{E},\beta,E_F}(E)$ given in Theorem 2.4.3 breaks down since $H_1(E)$ degenerates to mapping the entire left half-plane $\{E \mid \mathrm{Re}(E) < 0\}$ to 0. In this limit, we therefore replace $H_1(E)$ with the map $\tilde{H}_1(E) := E^2$ proposed in [LLYE09], which maps the imaginary axis onto $(-\infty, 0]$ and both $[-1,-\varepsilon]$, $[\varepsilon,1]$ to $[\varepsilon^2,1]$. The parameters of the map $H_2(x)$ are thus given by $a = \varepsilon^2$ and $b = 1$ in this case, from which the claim follows by repeating the above argument. □

2.5 Rational Interpolation vs. Contour Quadrature

Rational approximations are frequently constructed by discretising the contour integral in Cauchy's formula, i.e. by determining quadrature nodes z_k and weights λ_k such that

$$f(x) = \frac{1}{2\pi i} \int_{\partial\Omega} \frac{f(z)}{z - x}\, dz \approx \sum_{k=1}^{n} \lambda_k \, \frac{f(z_k)}{z_k - x} =: r_n(x), \tag{2.27}$$

where $\Omega \subset \mathbb{C}$ denotes some suitable contour integral domain. In [HHT08], such quadrature rules were constructed by mapping the above integral to an annulus and applying the periodic trapezoidal rule, and this idea has been applied to the Fermi-Dirac function f_{β,E_F} in [LLYE09]. This section will show that the contour quadrature ansatz described above is closely related to the rational interpolation scheme presented in previous sections but performs worse by a factor of two compared to our approach.

In the notation of Problem 2.0.1 and assuming for simplicity that both $\mathcal{E}$ and $\mathcal{S}$ are intervals or rays, the scheme from [HHT08] consists of the following steps.

1. Find an analytic, bounded and invertible map

$$\phi : \{z \mid \rho^{-1} < |z| < \rho\} \to \mathbb{C} \setminus \big(\mathcal{E} \cup \mathcal{S}\big)$$

   for some $\rho \in (0, \infty)$ determined by $\mathcal{E}$ and $\mathcal{S}$.

2. Substitute $x = \phi(z)$ in the contour integral from (2.27) and apply the trapezoidal rule,

$$f(x) = \frac{1}{2\pi i} \int_{|z|=1} \frac{f\big(\phi(z)\big)}{\phi(z) - x}\, \phi'(z)\, d\hat{z} \approx \frac{1}{n} \sum_{k=1}^{n} \frac{f\big(\phi(z_k)\big)}{\phi(z_k) - x}\, \phi'(z_k)\, z_k =: r_n(x)$$

   where $z_k = e^{2\pi i\, k/n}$ and the contour domain Ω in (2.27) is assumed to be given by $\Omega = \phi\big(\{|z| = 1\}\big)$.

It follows from standard convergence estimates for the periodic trapezoidal rule that the approximations $r_n(x)$ produced by this scheme satisfy the error bound

$$\|f - r_n\| \lesssim_\varepsilon \rho^{-n}, \tag{2.28}$$

see [TW14, Theorem 2.2].

To see the connection with rational interpolation, we note that if we compose the map $\phi(z)$ with an exponential, we obtain a function $\phi\big(\exp(z)\big)$ which instead of an annulus maps from the rectangle

$$\log\bigl(\{z \mid \rho^{-1} < |z| < \rho\}\bigr) = \bigl\{z \mid \mathrm{Re}(z) \in [-\log(\rho), \log(\rho)],\, \mathrm{Im}(z) \in (-\pi, \pi]\bigr\},$$

i.e. $\phi\bigl(\exp(z)\bigr)$ performs exactly the opposite mapping of the log-map $G_{\mathcal{E},\beta,E_F}(E)$ illustrated in Figure 2.6. Comparing the side lengths of these rectangles, we conclude $\log(\rho) = V(\mathcal{E}, \beta, E_F)/2$ and hence the convergence rate of the scheme from [LLYE09] is indeed half of ours as claimed above. This is illustrated in Figure 2.7, where we note that the x-axis in Figure 2.7a spans half the range of that of Figure 2.7b.

2.6 Rational Interpolation and Zolotarev Functions

It is known that for every real, continuous function $f(x)$ there exists a unique best real rational approximation

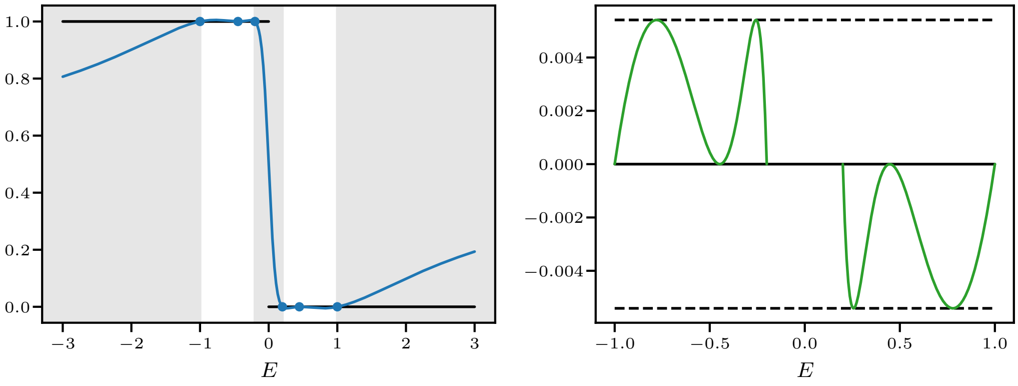
$$r^\star(x) := \operatorname*{arg\,min}_{r \in \mathcal{R}_{mn}^{\mathrm{real}}} \|f - r\|_{[-1,1]} \quad \text{where} \quad \mathcal{R}_{mn}^{\mathrm{real}} := \bigl\{r \in \mathcal{R}_{mn} \mid r\bigl([-1,1]\bigr) \subset \mathbb{R}\bigr\},$$

and this best approximation is characterised by a certain equioscillation property of the error $r^\star(x) - f(x)$, see [Tre13, Theorem 24.1]. Such best approximations must generally be computed using some iterative scheme like the Remez algorithm (see e.g. [Bra86, §V.6.B]) or the method proposed in [Mou16], but there is an important special case where the best rational approximation problem can be solved explicitly, namely the approximation of the sign function on domains of the form $\mathcal{E} = [-1, -\varepsilon] \cup [\varepsilon, 1]$ with $\varepsilon \in (0, 1)$ as shown by Zolotarev in 1877, see [Zol77]. This section will demonstrate that rational interpolation as discussed above essentially reproduces Zolotarev's solution if used appropriately.

Up to a linear transformation, the problem considered by Zolotarev is equivalent to the approximation of the zero-temperature Fermi-Dirac function

$$f_{\infty,0}(E) = \tfrac{1}{2}\bigl(1 - \mathrm{sign}(\mathrm{Re}(x))\bigr)$$

on the symmetric insulator-spectrum $\mathcal{E} = [-1, -\varepsilon] \cup [\varepsilon, 1]$. We have seen in the proof of (2.25) that the log-map $G_{\mathcal{E},\beta,E_F}(E)$ from Theorem 2.4.3 breaks down in the zero-temperature limit but that our construction of $G_{\mathcal{E},\beta,E_F}(E)$ can be adapted to this case if we replace the map $H_1(E)$ given in Theorem 2.4.3 with $\tilde{H}_1(E) := E^2$ and modify $H_2(E)$ accordingly. This gives us a new log-map $\tilde{G}_\varepsilon(E) \propto \tilde{H}_2\bigl(\tilde{H}_1(E)\bigr)$ which maps the upper-right quadrant $\bigl\{E \mid \mathrm{Re}(E) > 0, \mathrm{Im}(E) > 0\bigr\}$ holomorphically onto the rectangle $\bigl\{z \mid \mathrm{Re}(z) \in [0, V], \mathrm{Im}(z) \in [0, \pi]\bigr\}$ for some $V > 0$, and $\tilde{G}_\varepsilon\bigl([\varepsilon, 1]\bigr) =$



(a) Interpolant $r(E)$

(b) Error $r(E) - f_{\infty,0}(E)$

Figure 2.11: Rational interpolant $r(E) \in \mathcal{R}_{4,4}$ to the zero-temperature Fermi-Dirac function $f_{\infty,0}(E)$ on $\mathcal{E} := [-1,-0.2] \cup [0.2,1]$ (left) and the error $r(E) - f_{\infty,0}(E)$ (right). See Section 2.6 for further details.

$[0,\pi i]$, $\tilde G_\varepsilon\big([0,\infty i)\big) = [V, V+\pi i]$. Let us now introduce the interpolation points

$$X := X' \cup \big(-X'\big) \qquad \text{with} \qquad X' := \big\{\tilde G_\varepsilon^{-1}\big(\tfrac{\pi i k}{n}\big) \mid k = 0,\ldots,n\big\} \subset [\varepsilon, 1]$$

and poles

$$Y := Y' \cup \big(-Y'\big) \quad \text{with} \quad Y' := \big\{\tilde G_\varepsilon^{-1}\big(V + \tfrac{\pi i\,(k-1/2)}{n}\big) \mid k = 1,\ldots,n\big\} \subset (0,\infty i),$$

and define r as the rational interpolant to $f_{\infty,0}(E)$ in the sense of Theorem 2.1.1. We numerically observe that the resulting interpolant is in $\mathcal{R}_{2n,2n}$ even though our construction only guarantees $r \in \mathcal{R}_{2n+1,2n}$, and Figure 2.11b further demonstrates that the error function $r(E) - f_{\infty,0}(E)$ equioscillates on each of the two intervals $[-1,-\varepsilon]$ and $[\varepsilon,1]$, i.e. it assumes the extrema 0 and $\pm\|r(E) - f_{\infty,0}(E)\|_{[\varepsilon,1]}$ in $2n+1$ points in each interval. These observations combined with the equioscillation theorem from [Tre13, Theorem 24.1] imply that $r(E)$ is the best rational approximation to $f_{\infty,0}(E)$ up to a linear transformation, or more precisely

$$\tfrac{1}{1+e}\,\big(r(E) - \tfrac12\big) + \tfrac12 = \operatorname*{arg\,min}_{r \in \mathcal{R}^{\text{real}}_{2n,2n}} \|f_{\infty,0} - r\|_{[-1,1]}$$

where $e := \|r(E) - f_{\infty,0}(E)\|_{[\varepsilon,1]}$.

A very similar connection between holomorphic maps and Zolotarev's functions has previously been observed in [HHT08] in the context of contour quadrature. Furthermore, it has been shown in [NF16] that Zolotarev functions of dense matrices can be evaluated very efficiently by composing Zolotarev functions of lower degrees. However, we believe that this evaluation-by-composition scheme loses its efficiency if the argument is a sparse rather than a dense matrix, and hence it is unclear whether the observations from [NF16] can be applied in electronic structure algorithms.

2.7 Conclusion

This chapter has highlighted the connections between logarithmic potential theory, polynomial and rational approximation, and electronic structure theory, which culminated in explicit formulae for the rate of convergence of polynomial approximation to the Fermi-Dirac function in Theorem 2.3.9 and a novel rational interpolation scheme in Section 2.4. The formulae from Theorem 2.3.9 and in particular the asymptotics from Theorem 2.3.12 are useful for estimating the costs of electronic structure algorithms based on polynomial approximation (cf. Section 1.3) and for estimating the localisation of the density matrix as we will demonstrate in Theorem 3.2.2 in the next chapter. The rational interpolation scheme from Section 2.4 has been shown to converge faster than the one from [LLYE09] by a factor of two, but it loses out to the scheme from [Mou16] by a factor of two. The main contribution of our scheme is hence not superior approximation power, but rather the fact that it provides a generic framework for thinking about rational approximations which may be valuable in a number of future extensions.

To illustrate the last point, we would like to mention that it is occasionally of interest to evaluate the density matrix $f_{\beta,E_F}(H)$ not just for single values of β and E_F but rather for ranges of these parameters, e.g. for determining the Fermi level E_F by applying a root-finding algorithm to (1.6). The theory presented in this chapter demonstrates that we can achieve uniform accuracy for a range of β by choosing the set of poles Y according to the largest β since we have $\mathcal{S}^{\text{filled}}_{\beta,E_F} \subset \mathcal{S}^{\text{filled}}_{\beta',E_F}$ for $\beta < \beta'$, but the same does not hold if we sweep across a range of Fermi levels $E_F \in [E_F^{\min}, E_F^{\max}]$: to achieve uniform accuracy in the latter case, we should replace $\mathcal{S}^{\text{filled}}_{\beta,E_F}$ with the rectangular region

$$\mathcal{S}' := \Big\{ z \;\Big|\; \operatorname{Re}(z) \in [E_F^{\min}, E_F^{\max}],\ |\operatorname{Im}(z)| > \tfrac{\pi}{\beta} \Big\}$$

and interpolate with poles distributed along the boundary of $\mathcal{S}'$ with density given by $\mu_{\mathcal{E},\mathcal{S}',1}$.

Of course, the practical implementation of our rational interpolation scheme requires that we are able to solve the equilibrium measure problem, and the Schwarz-

Christoffel framework employed in this chapter occasionally requires manual fine-tuning or may even fail as we have seen in Theorem 2.4.5. A numerical method which allows us to solve equilibrium measure problems in a more black-box manner has been proposed in [Olv11], and we intend to investigate its application to rational approximation in future work.

Chapter 3

Incomplete Selected Inversion

This chapter tackles the second of the PEXSI subproblems introduced in Section 1.4, which is the fast evaluation of selected entries of the inverse. We have seen in Table 1.1 that the selected inversion algorithm from [ET75] allows us to compute these entries at reduced costs compared to evaluating the full inverse, but the costs scale worse than $\mathcal{O}(m)$ in dimensions $d > 1$. Aiming to overcome this limitation, we will show in Theorem 3.2.5 that the triangular factorisation computed as part of the selected inversion algorithm exhibits a localisation property similar to that of the density matrix discussed in Section 1.3, and we will propose in Section 3.3 an incomplete selected inversion algorithm which exploits this property to reduce costs to $\mathcal{O}(m)$ in all dimensions. In order to prepare for these developments, we begin this chapter with a brief review of the exact selected inversion algorithm in Section 3.1.

The triangular factorisation part of the incomplete selected inversion algorithm presented in Section 3.3 is exactly the symmetric version of the incomplete LU factorisation commonly used as a preconditioner in iterative methods for linear systems, see e.g. [Saa03, §10.3], and in addition to establishing a linear-scaling rational electronic structure algorithm, our analysis sheds a new light on this well-known algorithm.

3.1 Review of the Exact Selected Inversion Algorithm

We recall from Section 1.4 that the selected inversion algorithm applied to a matrix A consists in first computing a triangular factorisation of A and then inferring the values $A^{-1}(i,j)$ from this factorisation. This section will introduce the appropriate triangular factorisation in Subsection 3.1.1, discuss the cost of computing it for sparse matrices A in Subsection 3.1.2, and finally describe how to compute selected entries of the inverse in Subsection 3.1.3.

3.1.1 LDL^T Factorisation

As discussed in Section 1.3, the Hamiltonian matrices H considered in this thesis are real and symmetric, but due to the shifts z_k introduced in (1.16), the matrices $A = H - z_k$ to factorise are complex symmetric, i.e. $A \in \mathbb{C}^{m \times m}$ and $A^T = A$ but not $A^H = A$. The appropriate triangular factorisation for such matrices is the LDL^T factorisation introduced in the following theorem.

Theorem 3.1.1 ([GV96, Theorem 3.2.1]) *Let $A \in \mathbb{C}^{m \times m}$ be a symmetric matrix ($A = A^T$) such that all the leading submatrices $A(\ell, \ell)$ with $\ell = \{1, \dots, i\}$ and i ranging from 1 to m are invertible. Then, there exist matrices $L, D \in \mathbb{C}^{m \times m}$ such that L is lower-triangular with unit diagonal, D is diagonal and $A = LDL^T$.*

Definition 3.1.2 *For the remainder of this section, we let $A \in \mathbb{C}^{m \times m}$ be a symmetric matrix satisfying the conditions of Theorem 3.1.1, and we denote its LDL^T factors by $L, D \in \mathbb{C}^{m \times m}$. The indices $i, j \in \{1, \dots, m\}$ refer to an entry in the lower triangle, i.e. $i \geq j$, and we set $\ell := \{1, \dots, j-1\}$, $\bar{\ell} := \{1, \dots, j\}$, $r := \{j+1, \dots, m\}$ and $\bar{r} := \{j, \dots m\}$.*

The LDL^T factorisation of a given matrix A may be computed using the well-known Gaussian elimination algorithm [GV96, §3.2] which we will derive from the following result.[1]

Theorem 3.1.3 *In the notation of Definition 3.1.2, we have that*

$$L(i,j)\, D(j,j) = A(i,j) - L(i,\ell)\, D(\ell,\ell)\, L^T(\ell,j). \tag{3.1}$$

We observe that the right-hand side of (3.1) depends only on entries $L(i,k)$, $D(k,k)$ with $k < j$, hence the two factors can be computed by starting with $D(1,1) = A(1,1)$, $L(i,1) = A(i,1)/D(i,1)$ and proceeding iteratively in left-to-right order as follows.

Algorithm 3.1 LDL^T factorisation

```
for $j = 1, \dots, m$ do
    $D(j,j) = A(j,j) - L(j,\ell)\, D(\ell,\ell)\, L^T(\ell,j)$
    $L(r,j) = \big(A(r,j) - L(r,\ell)\, D(\ell,\ell)\, L^T(r,j)\big)/D(j,j)$
end for
```

[1] Both Theorem 3.1.3 and Lemma 3.1.4 were derived independently by the author, but given the importance of triangular factorisations and the simplicity of our formulae, we assume that similar statements have appeared previously in the literature.

We will show Theorem 3.1.3 and several other results in this chapter using the following auxiliary result.

Lemma 3.1.4 *In the notation of Definition* 3.1.2*, we have that*

$$L(i,j)\,D(j,j) = A(i,j) - A(i,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,j). \tag{3.2}$$

Proof. The matrices

$$\hat{L} := \begin{pmatrix} \mathbb{I} & \\ A(\bar{r},\ell)A(\ell,\ell)^{-1} & \mathbb{I} \end{pmatrix}, \qquad \hat{D} := \begin{pmatrix} A(\ell,\ell) & \\ & S \end{pmatrix} \tag{3.3}$$

with

$$S := A(\bar{r},\bar{r}) - A(\bar{r},\ell)\,A(\ell,\ell)^{-1}\,A(\ell,\bar{r})$$

provide a block LDL^T factorisation of A from which the full factorisation follows by further factorising

$$L_\ell D_\ell L_\ell^T := A(\ell,\ell), \qquad L_{\bar{r}} D_{\bar{r}} L_{\bar{r}}^T := S$$

and setting

$$L = \hat{L}\begin{pmatrix} L_\ell & \\ & L_{\bar{r}} \end{pmatrix}, \qquad D = \begin{pmatrix} D_\ell & \\ & D_{\bar{r}} \end{pmatrix}.$$

We thus compute

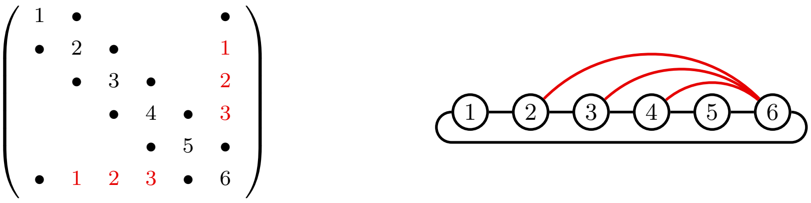
$$\begin{aligned} L(i,j)\,D(j,j) &= L_{\bar{r}}(i,j)\,D_{\bar{r}}(j,j) = L_{\bar{r}}(i,j)\,D_{\bar{r}}(j,j)L_{\bar{r}}^T(j,j) \\ &= S(i,j) = A(i,j) - A(i,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,j), \end{aligned}$$

where we enumerated the rows and columns of $L_{\bar{r}}, D_{\bar{r}}$ starting from j rather than 1 for consistency with the indexing in the full matrices. □

Proof of Theorem 3.1.3. It follows from the special structure of L and D that

$$\begin{aligned} A(i,\ell) &= L(i,\ell)\,D(\ell,\ell)\,L^T(\ell,\ell), \\ A(\ell,\ell)^{-1} &= L(\ell,\ell)^{-T}\,D(\ell,\ell)\,L(\ell,\ell)^{-1}, \\ A(\ell,j) &= L(\ell,\ell)\,D(\ell,\ell)\,L^T(\ell,j). \end{aligned}$$

The claim follows by inserting these expressions into (3.2). □



(a) Matrix

(b) Graph

Figure 3.1: Illustration of fill-in and level-of-fill for a one-dimensional periodic chain. The black numbers on the diagonal enumerate the vertices, the black dots indicate nonzero off-diagonal elements of the matrix, and the red numbers show the level-of-fill.

3.1.2 Sparse Factorisation and Nested Dissection

If A is sparse, its L-factor will generally have more nonzero entries than A, i.e. $\#\mathrm{fnz}(A) \geq \#\mathrm{nz}(A)$ in the notation from (1.18) and (1.17), respectively. The fill-in entries $(i,j) \in \mathrm{fnz}(A) \setminus \mathrm{fn}(A)$ significantly increase both the memory footprint of the factorisation and the cost of computing with it, hence it is important to reduce their number as much as possible. We recall the following theory from the literature (see e.g. [Dav06]) regarding the fill-in in sparse factorisations.

Definition 3.1.5 *The* graph $G(A) := \big(V(A), E(A)\big)$ *of a sparse matrix* $A \in \mathbb{C}^{m\times m}$ *is given by* $V(A) := \{1,\dots,m\}$ *and* $E(A) := \{(j,i) \mid A(i,j) \neq 0\}$*.*

Definition 3.1.6 *A* fill path *between two vertices* $i,j \in V(A)$ *is a path* $i, k_1, \dots, k_p, j$ *in* $G(A)$ *such that* $k_1,\dots,k_p < \min\{i,j\}$*.*

Theorem 3.1.7 ([RT78, Theorem 1]) *In the notation of Definition* 3.1.2 *and barring cancellation, we have that* $(i,j) \in \mathrm{fnz}(A)$ *if and only if there is a fill path between* i *and* j *in* $G(A)$*.*

Proof. The inverse $A(\ell,\ell)^{-1}$ from Lemma 3.1.4 can be written as a polynomial in $A(\ell,\ell)$, which shows $\big(A(\ell,\ell)^{-1}\big)\,(k_1,k_p) \neq 0$ if and only if $k_1, k_p \in \ell$ are connected in $G\big(A(\ell,\ell)\big)$. The term $A(i,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,1)$ hence yields $L(i,j) \neq 0$ if and only if i,j are connected by a fill path of length greater than 1, while the term $A(i,j)$ adds the entries corresponding to fill paths of length 1. □

Example 3.1.8 Consider a matrix with sparsity structure as shown in Figure 3.1. By Theorem 3.1.7, we get fill-in between vertices 4 and 6 because we can connect these two vertices via 3, 2 and 1 which are all numbered less than 4 and 6. We do not get fill-in between vertices 3 and 5, however, because all paths between these vertices have to go through either 4 or 6 which are larger than 3.

It follows from Theorem 3.1.7 that the number of fill-in entries depends not only on the sparsity pattern of A but also on the ordering. While finding an optimal fill-reducing ordering is an NP-hard problem [Yan81], the following nested dissection algorithm was shown to be asymptotically optimal up to at most a logarithmic factor in [Gil88].

Algorithm 3.2 Nested dissection

```
1: Partition the vertices into three sets $V_1, V_2, V_{sep}$ such that every path from $V_1$
   to $V_2$ visits at least one vertex in $V_{sep}$.
2: Arrange the vertices in the order $V_1, V_2, V_{sep}$, where $V_1$ and $V_2$ are ordered re-
   cursively according to the nested dissection algorithm and the ordering in $V_{sep}$
   is arbitrary.
```

The rationale for sorting the separator V_{sep} last on line 2 is that this eliminates all fill paths between V_1 and V_2 and thus $L(V_2, V_1) = 0$. The submatrix $L(V_{sep}, V_{sep})$ associated with the separator is typically dense, however, hence the nested dissection ordering is most effective if V_{sep} is small and V_1, V_2 are of roughly equal size.

The application of the nested dissection algorithm to a structured 2D mesh is illustrated in Figure 3.2. We note that the largest separator V_{sep} returned by this algorithm (the blue vertex set in the centre of Figure 3.2) contains $\mathcal{O}\big(\sqrt{m}\big)$ vertices; thus computing the associated dense part $L(V_{sep}, V_{sep})$ alone requires $\mathcal{O}\big(m^{3/2}\big)$ floating-point operations and the full factorisation must be at least as expensive to compute. It was shown in [Geo73] that this lower bound is indeed achieved, which justifies the ($d = 2$, Runtime) entry in Table 1.1 for the factorisation part of the selected inversion algorithm. The other entries can be derived along similar lines, see e.g. [Dav06], and the cost of the inversion part will be analysed in Theorem 3.1.12.

3.1.3 Selected Inversion

Given the LDL^T factorisation of a matrix A, its inverse can be computed using the following result.

Theorem 3.1.9 ([TFC73]) *In the notation of Definition* 3.1.2*, we have that*

$$A^{-1}(i,j) = D(i,j)^{-1} - A^{-1}(i,r)\, L(r,j). \tag{3.4}$$

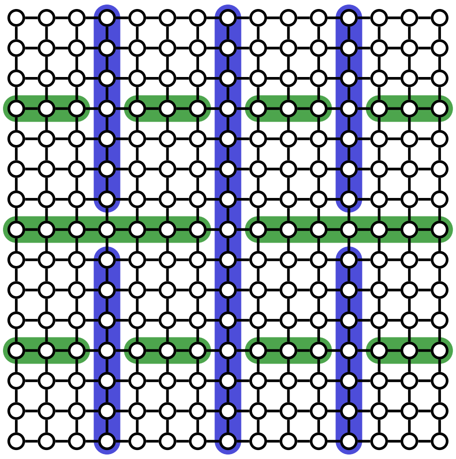


Figure 3.2: Nested dissection ordering of a structured 2D mesh. The vertices marked in blue and green denote alternating separators V_{sep}.

Proof. The claim follows from $A^{-1} = L^{-T}\,D^{-1} + A^{-1}\,(\mathbb{I} - L)$ which can be verified by substituting A^{-1} with $L^{-T}D^{-1}L^{-1}$. □

Equation (3.4) has the reverse property of (3.2): the right-hand side of (3.4) depends only on L, D and entries $A^{-1}(i,k)$ with $k > j$. The full inverse can thus be computed by starting with $A^{-1}(m,m) = D(m,m)^{-1}$ and iteratively growing the set of known entries in right-to-left order, but since the inverse of a sparse matrix is generally dense this would require at least $\mathcal{O}\big(m^2\big)$ floating-point operations. As noted in Section 1.4, we only need the entries $A^{-1}(i,j)$ with $(i,j) \in \operatorname{nz}(A)$ in electronic structure calculations, hence the question arises whether we can reduce the cost by computing only a subsets of the entries of A^{-1}. It was shown in [ET75] that this is indeed possible, and the following algorithm with

$$r^\circ = r^\circ(j) := \big\{ i \in \{j+1, \ldots, m\} \mid L(i,j) \neq 0 \big\} \tag{3.5}$$

was proposed to achieve this.

Algorithm 3.3 Selected inversion

```
for $j = m, \ldots, 1$ do
    $A^{-1}(r^\circ, j) = -A^{-1}(r^\circ, r^\circ)\, L(r^\circ, j)$
    $A^{-1}(j, r^\circ) = A^{-1}(r^\circ, j)^T$
    $A^{-1}(j,j) = D(j,j)^{-1} - A^{-1}(j, r^\circ)\, L(r^\circ, j)$
end for
```

Theorem 3.1.10 ([ET75]) *Algorithm 3.3 is correct, i.e. the computed entries*

$A^{-1}(i,j)$ agree with those of the exact inverse, and it is closed in the sense that all entries of A^{-1} required at iteration j have been computed in previous iterations $j' > j$.

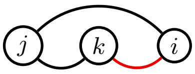
Proof. (Correctness.) The formulae in Algorithm 3.3 agree with those of Theorem 3.1.9 except that r° is used instead of r in the products $A^{-1}(i, r^\circ)\, L(r^\circ, j)$. This does not change the result of the computations since $L(r \setminus r^\circ, j) = 0$ by the definition of r°, hence the computed entries are correct.

(Closedness.) The entries of A^{-1} required on lines 3 and 4 are computed on line 2; thus it remains to show that the entries $A^{-1}(i,k)$ with $i, k \in r^\circ(j)$ required on line 2 have been computed in iterations $j' > j$. Due to the symmetry of A, we can assume $i \le k$ without loss of generality, and since the diagonal entry $A^{-1}(k,k)$ is explicitly computed on line 4 in iteration $j = k$, we may further restrict our attention to indices $i < k$. Such an entry $A^{-1}(i,k)$ is computed in iteration $j = k$ if and only if $i \in r^\circ(k)$, hence the claim follows from the following lemma. □

Lemma 3.1.11 ([ET75]) *In the notation of Definition* 3.1.2 *and with $r^\circ(j)$ as in* (3.5)*, we have that*

$$i, k \in r^\circ(j) \quad \text{and} \quad i > k \qquad \Longrightarrow \qquad i \in r^\circ(k).$$

Proof. According to Theorem 3.1.7, $i, k \in r^\circ(j)$ holds if and only if there exist fill paths from i and k to j, i.e. the graph structure is given by



where the two black edges indicate the aforementioned fill paths. Concatenating these two paths yields the red fill path from i to k and the claim follows. □

We note that Algorithm 3.3 computes exactly the entries $A^{-1}(i,j)$ with indices $(i,j) \in \operatorname{fnz}(A)$, hence its memory footprint is the same as that of the sparse factorisation. The following result establishes that the operation counts are also asymptotically the same.

Theorem 3.1.12 *The numbers of floating-point operations required by Algorithm* 3.1 *(sparse factorisation) and Algorithm* 3.3 *(selected inversion) are within a constant factor of each other.*

Proof. Algorithm 3.1 consists of the two operations

- $c(i,j) - L(i,k)\, D(k,k)\, L^T(k,j)$, performed if and only if $k < j \leq i$ and there are fill paths from j to k and k to i, and
- $c(i,j)/D(j,j)$, performed if and only if $j < i$ and there exists a fill path from j to i,

where in both cases $c(i,j)$ denotes an unspecified temporary variable. Similarly, Algorithm 3.3 consists of the two steps

- $c(i,k) - A^{-1}(i,j)\, L(j,k)$ and $c(i,k) - A^{-1}(j,i)\, L(i,k)$, performed if and only if $k < j \leq i$ and there are fill paths from k to j and k to i (only one of the two operations is performed if $i = j$), and
- $c(j,j) - A^{-1}(j,i)\, L(i,j)$, performed if and only if $j < i$ and there exists a fill path from j to i.

The claim follows by noting that both algorithms perform $\mathcal{O}(1)$ operations for each triplet (k,i,j) and pair (i,j) satisfying the same conditions. □

3.2 Exponential Localisation

This section will establish in Theorem 3.2.5 below that the LDL^T factorisation required by the selected inversion algorithm exhibits a localisation property similar to that of the density matrix $f_{\beta,E_F}(H)$ described in (1.10). To prepare for this result, we first recall in Theorem 3.2.2 a precise formulation regarding the localisation of the inverses $(H-z)^{-1}$ from [DMS84].

Definition 3.2.1 *We use the following notation throughout this section.*

- $H \in \mathbb{C}^{m\times m}$ *denotes a sparse, symmetric matrix, and* $i,j \in \{1,\dots,m\}$ *are indices for an entry in the lower triangle, i.e.* $i \geq j$.
- $\mathcal{E} \subset \mathbb{R}$ *denotes a non-polar set (cf. Definition* 2.1.5*) such that the spectra of all leading submatrices* $H(\ell,\ell)$ *with* $\ell = \{1,\dots,i\}$ *and* i *ranging from* 1 *to* m *are contained in* $\mathcal{E}$. *We will further comment on this assumption in Remark* 3.2.6.
- $z \in \mathbb{C} \setminus \mathcal{E}$ *denotes a point oustide* $\mathcal{E}$.
- $g_{\mathcal{E}}(z)$ *denotes the Green's function associated with* $\mathcal{E}$ *(see Definition* 2.3.1*).*
- L, D *denote the* LDL^T *factors of* $H - z$.
- $d(i,j)$ *denotes the graph distance in* $G(H)$*, which is defined as the minimal number of edges on any path between* i *and* j*, or* ∞ *if there are no such paths.*

Theorem 3.2.2 ([DMS84]) *In the notation of Definition* 3.2.1*, we have that*

$$\big|(H-z)^{-1}(i,j)\big| \lesssim_{\varepsilon} \exp\big(-g_{\mathcal{E}}(z)\, d(i,j)\big).$$

(The notation $\lesssim_\varepsilon$ was introduced in Definition 2.1.9.)

The proof of Theorem 3.2.2 follows immediately from Theorem 2.3.8 (convergence of polynomial approximation) and the following lemma.

Lemma 3.2.3 ([DMS84]) *In the notation of Definition 3.2.1, we have for all bounded $f : \mathcal{E} \to \mathbb{C}$ that*

$$\big|f(H)(i,j)\big| \le \inf_{p \in \mathcal{P}_{d(i,j)-1}} \|f - p\|_{\mathcal{E}}$$

where $\mathcal{P}_k$ denotes the space of polynomials of degree $\le k$.

Proof. Since $p \in \mathcal{P}_{d(i,j)-1}$, we have that $p(H)(i,j) = 0$ and thus

$$\begin{aligned} |f(H)(i,j)| &\le |p(H)(i,j)| + |f(H)(i,j) - p(H)(i,j)| \\ &\le 0 + \|f(H) - p(H)\|_2 \\ &\le \|f - p\|_{\mathcal{E}}. \end{aligned}$$

□

We next derive a localisation result for the LDL^T factorisation of $H - z$ by applying Theorem 3.2.2 to the inverse $A(\ell,\ell)^{-1}$ from Lemma 3.1.4. In order to formulate this result, we need a new notion of distance defined as follows.

Definition 3.2.4 ([Saa03, §10.3.3]) *In the notation of Definition 3.2.1, the level-of-fill* level(i,j) *are given by*

$$\text{level}(i,j) := \max\{0, d_{\text{fill}}(i,j) - 1\}$$

where $d_{\text{fill}}(i,j)$ denotes the minimal number of edges on any fill path between i and j, or ∞ if no such path exists.

An example for the level-of-fill is provided in Figure 3.1.

Theorem 3.2.5 *In the notation of Definition 3.2.1, we have that*

$$|L(i,j)| \lesssim_\varepsilon \exp\big(-g_{\mathcal{E}}(z)\,\text{level}(i,j)\big).$$

Proof. The claim is trivially true if level$(i,j) = 0$, hence we restrict ourselves to i,j such that level$(i,j) > 0$ and $H(i,j) = 0$ in the following. According to Lemma 3.1.4, we then have that

$$L(i,j) = -A\big(i, \ell^\circ(i)\big)\, A_\ell^{-1}\big(\ell^\circ(i), \ell^\circ(j)\big)\, A\big(\ell^\circ(j), j\big)/D(j,j)$$

where $A = H - z$, $A_\ell := A(\ell,\ell)$ with $\ell = \{1,\dots,j-1\}$, and $\ell^\circ(t) := \big\{k \in \ell \mid A(t,k) = A(k,t) \neq 0\big\}$. By the definition of level-of-fill, we have that $d_\ell(i^\circ,j^\circ) \geq \operatorname{level}(i^\circ,j^\circ) - 1$ for all $i^\circ \in \ell^\circ(i)$, $j^\circ \in \ell^\circ(j)$ with $d_\ell(i,j)$ the graph distance in $G(A_\ell)$, and thus $A_\ell(i^\circ,j^\circ) \lesssim_\varepsilon \exp\big(-g_{\mathcal{E}}(z)\operatorname{level}(i,j)\big)$ according to Theorem 3.2.2. The claim follows after noting that $\#\ell^\circ(i)$, $\#\ell^\circ(j)$ are bounded independently of m due to the sparsity of H, and that $|D(j,j)| \geq \min|z - \mathcal{E}|$ since $D(j,j)^{-1} = A_{\bar\ell}^{-1}(j,j)$ with $\bar\ell = \{1,\dots,j\}$ and the spectrum of $A_{\bar\ell}$ is contained in $\mathcal{E}$ according to the assumptions in Definition 3.2.1. □

Remark 3.2.6 The assumption that the spectra of all leading submatrices $H(\ell,\ell)$ are contained in $\mathcal{E}$ in Definition 3.2.1 was introduced specifically to allow for Theorem 3.2.5. We would like to point out that this assumption can always be satisfied by choosing $\mathcal{E}$ as the convex hull of the spectrum of H and that the rational approximation scheme constructed in Chapter 2 places the poles away from the real axis and hence outside of this convex hull. Furthermore, we expect that the spectra of the submatrices are usually contained in the spectrum of H since in physical terms this corresponds to the assumption that the electronic properties of subsystems agree with those of the overall system. If true, the extra condition in Definition 3.2.1 is redundant and we may choose $\mathcal{E}$ simply as the spectrum of H as in Definition 3.1.2. We will return to this point in Example 3.4.1.

We conclude from Theorems 3.2.2 and 3.2.5 that the entries of both the inverse $(H - z)^{-1}(i,j)$ and the L-factor $L(i,j)$ decay exponentially with the same rate $g_{\mathcal{E}}(z)$ but with different notions of distance $d(i,j)$ and $\operatorname{level}(i,j)$, respectively. This qualitative difference is illustrated in the following example.

Example 3.2.7 Consider the $m \times m$ matrix

$$A(i,j) := \begin{cases} 3 & \text{if } i = j \\ -1 & \text{if } i = j \pm 1 \mod m, \\ 0 & \text{otherwise} \end{cases}$$

whose graph structure for $m = 6$ is illustrated in Figure 3.1. We observe that $\operatorname{level}(m,j) = j - 1$ increases monotonically from $j = 1$ to $j = m - 2$ and thus $L(m,j)$ decreases monotonically over the same range, see Figure 3.3b. Conversely, $d(m,j) = \min\{j, m-j\}$ has a maximum at $j = \frac{m}{2}$ and thus $|A^{-1}(m,j)|$ has a minimum at this value of j, see Figure 3.3a.

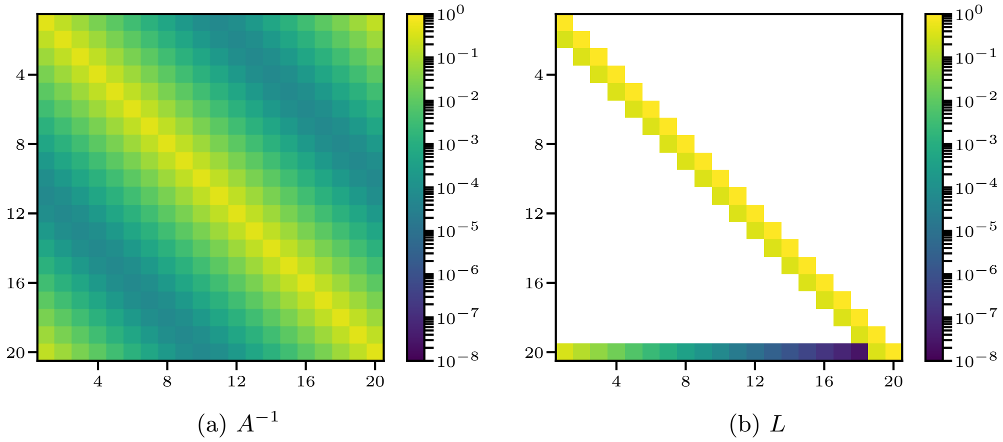


Figure 3.3: Decay in the inverse and L-factor of the matrix from Example 3.2.7.

3.3 Incomplete Selected Inversion

Theorem 3.2.5 asserts that entries $L(i,j)$ with a large level-of-fill $\operatorname{level}(i,j)$ are small, which raises the question whether compute time and memory can be saved in the sparse factorisation step of the selected inversion algorithm by only computing entries $L(i,j)$ with indices in the restricted set

$$\operatorname{ifnz}(H) = \big\{(i,j) \in \operatorname{fnz}(H) \mid \operatorname{level}(i,j) \le c\big\} \tag{3.6}$$

for some cut-off level-of-fill $c \ge 0$. Of course, ignoring entries which are small but nonzero introduces errors and the incomplete factorisation suggested above is only worth considering if the errors are small and the savings substantial.

We will see in Theorem 3.3.3 that restricting the sparse factorisation to (3.6) reduces the compute time and memory requirements to $\mathcal{O}(m)$ independently of the dimension, which is significantly lower than the costs of the exact algorithm listed in Table 1.1. Regarding the error, we will see in Theorem 3.3.4 that dropping entries in the LDL^T factorisation of a matrix A corresponds to computing the exact factorisation $\tilde{A} = \tilde{L}\tilde{D}\tilde{L}^T$ of a perturbed matrix $\tilde{A} \approx A$, and this will allows us to bound the error in the inverse $\|A^{-1} - \tilde{A}^{-1}\|_{\operatorname{nz}(A)}$ as a function of the cut-off level-of-fill c in Corollary 3.3.10 assuming a certain conjecture regarding the magnitudes of the dropped entries holds true. Finally, the following result translates the errors in the inverses into an error bound for the quantities of interest.

Theorem 3.3.1 *Consider a quantity of interest q and an approximation $\tilde{q}$ given,* *respectively, by*

$$q := \sum_{k=1}^{n} c_k \, \mathrm{Tr}\big(M \, A_k^{-1}\big) \qquad \text{and} \qquad \tilde{q} := \sum_{k=1}^{n} c_k \, \mathrm{Tr}\big(M \, \tilde{A}_k^{-1}\big),$$

where $c_k, z_k \in \mathbb{C}$ *and* $M, A_k, \tilde{A}_k \in \mathbb{C}^{m \times m}$ *with* $\#\mathrm{nz}(M) = \mathcal{O}(m)$. *We then have that*

$$|q - \tilde{q}| \leq \mathcal{O}(m) \sum_{k=1}^{n} |c_k| \, \|A_k^{-1} - \tilde{A}_k^{-1}\|_{\mathrm{nz}(M)}, \tag{3.7}$$

where $\|A\|_{\mathcal{I}}$ *denotes the maximum norm on* $\mathcal{I} \subset \{1, \ldots, m\}^2$,

$$\|A\|_{\mathcal{I}} := \max_{(i,j) \in \mathcal{I}} |A(i,j)|.$$

Proof. We compute

$$|q - \tilde{q}| = \left| \sum_{k=1}^{n} c_k \, \mathrm{Tr}\big(M \, (A_k^{-1} - \tilde{A}_k^{-1}\big) \right| \leq \sum_{k=1}^{n} |c_k| \sum_{i,j=1}^{m} |M(i,j)| \, \|A_k^{-1} - \tilde{A}_k^{-1}\|_{\mathrm{nz}(M)}$$

and note that $\sum_{i,j=1}^{m} |M(i,j)| = \mathcal{O}\big(\mathrm{nz}(M)\big) = \mathcal{O}(m)$. □

We remark that the $\mathcal{O}(m)$-factor in the bound (3.7) is acceptable in applications since accuracy is usually required on a per-atom basis. In conclusion, we have thus seen that restricting the factorisation to only the entries (3.6) leads to substantial computational savings and the error can be controlled by choosing the cut-off level-of-fill c large enough.

Dropping entries only in the factorisation step does not reduce the asymptotic costs of the selected inversion algorithm, however, since the selected inversion step must still compute all the entries $\tilde{A}^{-1}(i,j)$ with $(i,j) \in \mathrm{fnz}(A)$ even when applied to an incomplete factorisation $A \approx \tilde{L}\tilde{D}\tilde{L}^T$ in order to preserve the closedness property from Theorem 3.1.10. We address this issue in Subsection 3.3.2 by proposing an incomplete selected inversion algorithm which computes approximate entries $B(i,j) \approx \tilde{A}^{-1}(i,j)$ only for $(i,j) \in \mathrm{ifnz}(A)$. We will see in Theorem 3.3.11 that the computational complexity of this algorithm is $\mathcal{O}(m)$, and we will bound the error $\|\tilde{A}^{-1} - B\|_{\mathrm{nz}(A)}$ as a function of c in Corollary 3.3.15 again assuming a conjecture regarding the magnitudes of the dropped entries holds true. Using the triangle inequality, this then yields the total error bound

$$\|A^{-1} - B\|_{\mathrm{nz}(A)} \leq \|A^{-1} - \tilde{A}^{-1}\|_{\mathrm{nz}(A)} + \|\tilde{A}^{-1} - B\|_{\mathrm{nz}(A)}, \tag{3.8}$$

which can be made arbitrarily small by choosing c large enough.

Definition 3.3.2 *This section follows the notation of Definitions* 3.1.2 *and* 3.2.1 *with $A = H - z$ as well as the following additions.*

- $\tilde{r} := \{k \in r \mid (k, j) \in \operatorname{ifnz}(H)\}$ *with j and r as in Definition* 3.1.2.
- *c denotes the cut-off level-of-fill from* (3.6).
- *$\tilde{L}, \tilde{D}$ denote the incomplete LDL^T factors and E denotes the dropped entries computed in Algorithm* 3.4. *Furthermore, we set $\tilde{A} := \tilde{L}\tilde{D}\tilde{L}^T$.*
- *$B(i, j) \approx A^{-1}(i, j)$ denotes the approximate entries of the inverse and F denotes the dropped entries computed in Algorithm* 3.5.
- *In both Algorithms* 3.4 *and* 3.5, *we assume that matrix entries which are not specified are set to zero.*

3.3.1 Incomplete LDL^T Factorisation

We propose the following incomplete algorithm for the sparse factorisation step.

Algorithm 3.4 Incomplete LDL^T factorisation

```
1: for $j = 1, \dots, m$ do
2:     $\tilde{D}(j,j) = A(j,j) - \tilde{L}(j,\ell)\,\tilde{D}(\ell,\ell)\,\tilde{L}^T(\ell,j)$
3:     $\tilde{L}(\tilde{r},j) = \big(A(\tilde{r},j) - \tilde{L}(\tilde{r},\ell)\,\tilde{D}(\ell,\ell)\,\tilde{L}^T(\ell,j)\big)/\tilde{D}(j,j)$
4:     $E(r \setminus \tilde{r}, j) = \tilde{L}(r \setminus \tilde{r},\ell)\,\tilde{D}(\ell,\ell)\,\tilde{L}^T(\ell,j)$
5:     $E(j, r \setminus \tilde{r}) = E(r \setminus \tilde{r}, j)^T$
6: end for
```

We note that this is exactly the symmetric version of the well-known incomplete LU factorisation commonly used as a preconditioner in iterative methods for linear systems, see e.g. [Saa03, §10.3].

Theorem 3.3.3 *Algorithm* 3.4 *requires $\mathcal{O}(m)$ runtime and memory.*

Proof. For every fixed $j \in \{1, \dots, m\}$, there are at most $\mathcal{O}\big(c^d\big)$ vertices within a distance c from j, where d denotes the dimension of the system under consideration. Every column $\tilde{L}(\cdot, j)$ thus has at most $\mathcal{O}\big(c^d\big)$ nonzero entries, which shows that every iteration of the loop in Algorithm 3.4 requires at most $\mathcal{O}\big(c^{2d}\big)$ floating-point operations. □

Keeping track of the dropped entries E in lines 4 and 5 is not required in an actual implementation, but doing so in Algorithm 3.4 allows us to conveniently formulate the following results.

Theorem 3.3.4 ([Saa03, Proposition 10.4]) *In the notation of Definition* 3.3.2*, we have that* $\tilde{L}\tilde{D}\tilde{L}^T = A + E$.

Proof. We note that level$(i, j) > 0$ for all $i \in r \setminus \tilde{r}$ and hence $A(r \setminus \tilde{r}, j) = 0$, which yields

$$A(r \setminus \tilde{r}, j) + E(r \setminus \tilde{r}, j) - \tilde{L}(r \setminus \tilde{r}, \ell)\, \tilde{D}(\ell, \ell)\, \tilde{L}^T(\ell, j) = 0.$$

Since $E(\tilde{r}, j) = 0$, we can rewrite line 3 of Algorithm 3.4 as

$$\tilde{L}(r, j) = \big(A(r, j) + E(r, j) - \tilde{L}(r, \ell)\, \tilde{D}(\ell, \ell)\, \tilde{L}^T(r, j)\big)/\tilde{D}(j, j),$$

and similarly we can rewrite line 2 as

$$\tilde{D}(j, j) = A(j, j) + E(j, j) - \tilde{L}(j, \ell)\, \tilde{D}(\ell, \ell)\, \tilde{L}^T(\ell, j)$$

since $E(j, j) = 0$. These are precisely the recursion formulae of the exact LDL^T factorisation in Algorithm 3.1 applied to the matrix $A + E$. The claim follows after noting that $A + E$ is symmetric because we explicitly symmetrise E on line 5. □

Theorem 3.3.5 *In the notation of Definition* 3.3.2 *and assuming* $\|E\|_2 < \delta := \min |z - \mathcal{E}|$*, we have that*

$$\begin{aligned}\big|\big((H - z)^{-1} - (H + E - z)^{-1}\big)(i, j)\big| &\lesssim_\varepsilon \dots \\ &\lesssim_\varepsilon \sum_{\tilde{\imath},\tilde{\jmath}=1}^{m} \exp\Big(-g_{\mathcal{E}}(z)\, \big(d(i, \tilde{\imath}) + d(\tilde{\jmath}, j)\big)\Big)\, \big|E(\tilde{\imath}, \tilde{\jmath})\big| + \frac{\delta^{-2}\, \|E\|_2^2}{\delta - \|E\|_2}.\end{aligned} \tag{3.9}$$

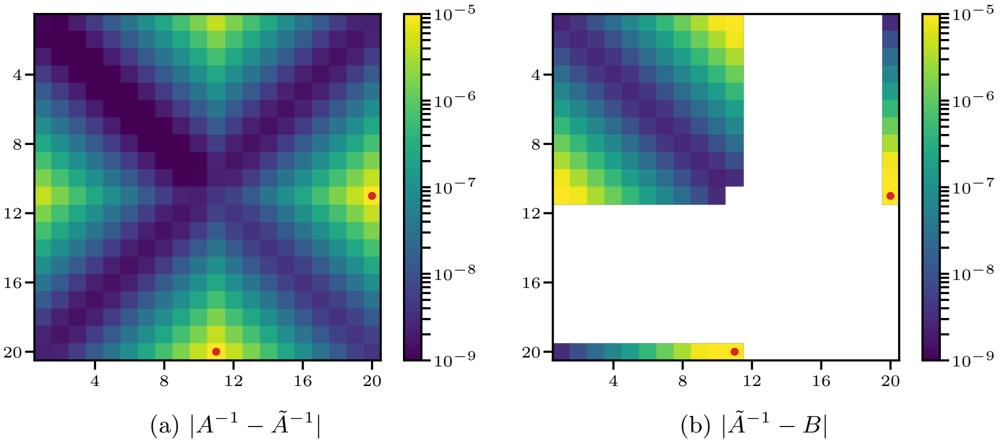
This bound is illustrated in Figure 3.4a.

Proof. Expanding $(H + E - z)^{-1}$ in a Neumann series around $H - z$, we obtain

$$\begin{aligned}&(H - z)^{-1} - (H + E - z)^{-1} = \dots \\ &\qquad = (H - z)^{-1}\, E\, (H - z)^{-1} - \sum_{k=2}^{\infty} \big(- (H - z)^{-1}\, E\big)^k\, (H - z)^{-1}.\end{aligned}$$

The claim follows by estimating the entries of $(H - z)^{-1}$ in the first term using Theorem 3.2.2 and bounding the entries of the second term through its operator norm. □

Theorem 3.3.5 provides an a-posteriori error estimate for the inverse $(H+E-z)^{-1}$ in terms of the dropped entries E, which could be used in an adaptive truncation



(a) $|A^{-1} - \tilde{A}^{-1}|$

(b) $|\tilde{A}^{-1} - B|$

Figure 3.4: Error introduced by the incomplete selected factorisation step (left) and incomplete selected inversion step (right) applied to the matrix from Example 3.2.7 with a cut-off level-of-fill $c = 9$. The red dots mark the nonzero entries in E and F, respectively.

scheme where $\operatorname{ifnz}(H)$ is of the form

$$\operatorname{ifnz}(H) = \big\{(i,j) \in \operatorname{fnz}(H) \mid |\tilde{L}(i,j)| \geq \tau\big\}$$

for some tolerance $\tau \geq 0$, see [Saa03, §10.4]. Conversely to the level-of-fill-based scheme from (3.6), such a tolerance-based scheme would control the error but not the amount of fill-in since the perturbed entries $|\tilde{L}(i,j)|$ may fail to be small even when the corresponding entries $L(i,j)$ are small. Both schemes thus require further information about the perturbed L-factor $\tilde{L}(i,j)$ or equivalently about the dropped entries $E(i,j)$ in order to simultaneously control the accuracy and the computational effort. More precisely, in the case of the level-of-fill scheme (3.6) we need to understand

- the sparsity pattern of E since this impacts the number of terms and the size of the exponential factor in (3.9), and
- the magnitudes of the nonzero entries $E(\tilde{\imath},\tilde{\jmath})$.

The first of these points is easily addressed.

Theorem 3.3.6 *In the notation of Definition* 3.3.2*, we have that*

$$E(i,j) \neq 0 \quad\implies\quad c < \operatorname{level}(i,j) \leq 2c+1.$$

In particular, the number of nonzero entries per row or column of E is bounded *independently of m.*

The proof of this result will make use of the following lemma.

Lemma 3.3.7 *In the notation of Definition* 3.3.2 *and barring cancellation, we have that $E(i,j) \neq 0$ if and only if $(i,j) \in \operatorname{fnz}(H) \setminus \operatorname{ifnz}(H)$ and there exists a $k \in \ell$ such that $(i,k),(j,k) \in \operatorname{ifnz}(H)$.*

Proof. The claim follows by noting that line 4 in Algorithm 3.4 performs a nonzero update on $E(i,j)$ if and only if

$$i \in r \setminus \tilde{r} \qquad \Longleftrightarrow \qquad (i,j) \in \operatorname{fnz}(H) \setminus \operatorname{ifnz}(H)$$

and there exists a $k \in \ell$ such that

$$\tilde{L}(i,k), \tilde{L}(j,k) \neq 0 \qquad \Longleftrightarrow \qquad (i,k),(j,k) \in \operatorname{ifnz}(H). \qquad \square$$

Proof of Theorem 3.3.6. According to Lemma 3.3.7, we have that

$$E(i,j) \neq 0 \implies (i,j) \in \operatorname{fnz}(H) \setminus \operatorname{ifnz}(H) \implies \operatorname{level}(i,j) > c.$$

To derive the upper bound, let us assume that $E(i,j) \neq 0$. Then, Lemma 3.3.7 guarantees that there exists a vertex $k \in \ell$ such that i,k and j,k are connected by fill paths of lengths at most $c+1$ (recall from Definition 3.2.4 that the level-of-fill is the length of the shortest path minus 1). Concatenating these two paths yields a fill path between i and j of length at most $2c+2$, hence $\operatorname{level}(i,j) \leq 2c+1$. Finally, the claim regarding the sparsity of E follows by noting that there are at most $\mathcal{O}\big(c^d\big)$ vertices i within a distance $2c+2$ from j, where d denotes the dimension of the system under consideration. $\square$

Estimating the magnitudes $|E(\tilde{\imath},\tilde{\jmath})|$ of the dropped entries proved to be challenging, and we have not managed to resolve this point conclusively. Our numerical experiments, to be presented in Section 3.4, suggest that a bound of the following form holds, and the subsequent discussion establishes the main obstacle which needs to be overcome in order to prove our claim.

Conjecture 3.3.8 *In the notation of Definition* 3.3.2*, we have that*

$$|E(i,j)| \lesssim_{\varepsilon} \exp\big(-g_{\mathcal{E}}(z)\operatorname{level}(i,j)\big).$$

Discussion. Reversing the substitutions in the proof of Theorem 3.1.3, we obtain

$$E(i,j) = \tilde{L}(i,\ell)\,\tilde{D}(\ell,\ell)\,\tilde{L}^T(\ell,j) = \tilde{A}(i,\ell)\,\tilde{A}(\ell,\ell)^{-1}\,\tilde{A}(\ell,j),$$

and expanding the latter formula to first order in $\|E\|_2$ as in Theorem 3.3.5 yields

$$\begin{aligned} E(i,j) = {} & A(i,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,j)\ldots \\ & + E(i,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,j)\ldots \\ & - A(i,\ell)\,A(\ell,\ell)^{-1}E(\ell,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,j)\ldots \\ & + A(i,\ell)\,A(\ell,\ell)^{-1}\,E(\ell,j) + \mathcal{O}\big(\|E\|_2^2\big). \end{aligned} \tag{3.10}$$

Theorem 3.2.5 guarantees that the first term

$$A(i,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,j) = L(i,j)\,D(j,j)$$

on the right-hand side of (3.10) satisfies $\big|L(i,j)\,D(j,j)\big| \lesssim_{\varepsilon} \exp\big(-g_{\mathcal{E}}(z)\,c\big)$, but bounding the remaining terms is challenging because the magnitudes of these terms recursively depend on the errors committed earlier.

To illustrate this point, let us assume we have a bound $|E(\tilde{\imath},\tilde{\jmath})| \leq C_0$ with $C_0 \sim \exp\big(-g_{\mathcal{E}}(z)\,c\big)$ for all entries of E on the right-hand side of (3.10) such that we can bound e.g. the second term by

$$\big|E(i,\ell)\,A(\ell,\ell)^{-1}\,A(\ell,j)\big| \leq C_0 \sum_{k\in\ell} \big|A_\ell^{-1}(k,\ell)\,A(\ell,j)\big| \tag{3.11}$$

where $A_\ell := A(\ell,\ell)$. From the sparsity of $A(\ell,j)$ and the localisation of A_ℓ^{-1}, it then follows that the sum on the right-hand side of (3.11) decays exponentially for an appropriate ordering of the terms and can therefore be bounded by some constant C independent of m. In general, this constant C will be larger than one, however, since some $k \in \ell$ may well be close to j in terms of the graph distance on $G(A_\ell)$ such that the corresponding terms are not small. Bounding the other terms in (3.10) similarly, we thus obtain $|E(i,j)| \leq C\,C_0$ for some constant $C > 1$.

For the next entry $E(i',j')$ to be estimated using (3.10), the entry $E(i,j)$ that we just estimated may now appear on the right-hand side such that we have to assume the bound $|E(\tilde{\imath}',\tilde{\jmath}')| \leq C\,C_0$ for these entries. Proceeding analogously as above, we then obtain the bound $|E(i',j')| \leq C^2\,C_0$ which is worse by a factor of $C > 1$ than the bound in the preceding step and worse by a factor of $C^2 > 1$ than the bound two steps ago. We therefore conclude that any estimate on the dropped entries $E(i,j)$ deteriorates exponentially with every recursive application of (3.10).

The key issue in the above analysis is that without further knowledge about the entries $E(i,j)$, we have to assume that all the error terms in the recursion formula (3.10) accumulate rather than cancel. We conjecture that such an accumulation of errors cannot occur, at least for matrices which are "well-behaved" in a suitable sense, but a rigorous proof of this claim requires deeper insight into the structure of the incomplete LDL^T factorisation and is left for future work. □

Conjecture 3.3.8 suggests that the incomplete factorisation exhibits the same localisation as the exact factorisation, and this is in principle enough to derive an a-priori bound from the a-posteriori bound (3.9). However, we introduce one more assumption in order to simplify the final result.

Assumption 3.3.9 *Either* $\operatorname{level}(i,j) = \infty$ *or* $\operatorname{level}(i,j) \sim d(i,j)$.

Discussion. We have seen in Example 3.2.7 that this assumption is not satisfied in the case of one-dimensional periodic chains, but we expect that this counterexample is the "exception which proves the rule". In particular, we conjecture that Assumption 3.3.9 is always satisfied in dimensions $d > 1$ and if the nested dissection ordering is used, since in this case every pair of vertices is connected by many paths and it seems unlikely that the nested dissection ordering would place a high-numbered vertex on all the short paths. This hypothesis is supported by our numerical experiments presented in Example 3.4.2 below. Furthermore, we will see in Example 3.4.3 that even if Assumption 3.3.9 is violated, the conclusions that we draw from it still hold up to some minor modifications. □

Corollary 3.3.10 *In the notation of Definition* 3.3.2*, and assuming Conjecture* 3.3.8 *and Assumption* 3.3.9*, we have that*

$$\big\|(H-z)^{-1} - (H+E-z)^{-1}\big\|_{\operatorname{nz}(H)} \lesssim_\varepsilon \exp\big(-2\, g_{\mathcal{E}}(z)\, c\big).$$

Proof. It follows from Theorem 3.3.6 (sparsity of E) and Conjecture 3.3.8 (localisation of E) that for all $\tilde\imath, \tilde\jmath \in \{1,\ldots,m\}$ we have that $|E(\tilde\imath,\tilde\jmath)| \lesssim_\varepsilon \exp\big(-g_{\mathcal{E}}(z)\, c\big)$, and inserting this estimate into the bound from Theorem 3.3.5 (a-posteriori error bound) yields

$$\begin{aligned} &\big|\big((H-z)^{-1} - (H+E-z)^{-1}\big)(i,j)\big| \lesssim_\varepsilon \ldots \\ &\qquad \lesssim_\varepsilon \sum_{(\tilde\imath,\tilde\jmath)\in\operatorname{nz}(E)} \exp\Big(-g_{\mathcal{E}}(z)\,\big(d(i,\tilde\imath) + d(\tilde\jmath,j) + c\big)\Big) + \frac{\delta^{-2}\,\|E\|_2^2}{\delta - \|E\|_2}. \end{aligned} \tag{3.12}$$

We are only interested in entries $(i,j) \in \mathrm{nz}(H)$ for which $d(i,j) \leq 1$; thus we conclude from the triangle inequality and Assumption 3.3.9 ($d(i,j) \sim \mathrm{level}(i,j)$) that for all $(\tilde\imath,\tilde\jmath) \in \mathrm{nz}(E)$ we have that

$$d(i,\tilde\imath) + d(\tilde\jmath,j) + 1 \geq d(\tilde\imath,\tilde\jmath) \sim \mathrm{level}(\tilde\imath,\tilde\jmath) \geq c+1.$$

In particular, we note that if $d(i,\tilde\imath) \lesssim \frac{c}{2}$, then $d(\tilde\jmath,j) \gtrsim \frac{c}{2}$ and vice versa, which allows us to bound the first term in (3.12) by

$$\begin{aligned}
&\sum_{(\tilde\imath,\tilde\jmath)\in\mathrm{nz}(E)} \exp\Big(-g_{\mathcal{E}}(z)\,\big(d(i,\tilde\imath) + d(\tilde\jmath,j) + c\big)\Big) \lesssim \dots \\
&\qquad \lesssim 2 \sum_{\substack{\tilde\imath \text{ such that} \\ d(i,\tilde\imath)\lesssim\frac{c}{2}}} \; \sum_{\substack{\tilde\jmath \text{ such that} \\ d(\tilde\jmath,j)\gtrsim c-d(i,\tilde\imath)}} \exp\Big(-g_{\mathcal{E}}(z)\,\big(d(i,\tilde\imath) + d(\tilde\jmath,j) + c\big)\Big) + \dots \\
&\qquad\qquad + \sum_{\substack{\tilde\imath \text{ such that} \\ d(i,\tilde\imath)\gtrsim\frac{c}{2}}} \; \sum_{\substack{\tilde\jmath \text{ such that} \\ d(\tilde\jmath,j)\gtrsim\frac{c}{2}}} \exp\Big(-g_{\mathcal{E}}(z)\,\big(d(i,\tilde\imath) + d(\tilde\jmath,j) + c\big)\Big) \\
&\qquad \lesssim_{\varepsilon} \exp\big(-2\,g_{\mathcal{E}}(z)\,c\big)
\end{aligned}$$

where on the last line we estimated the infinite sums using the boundedness of the geometric series and the finite sum over $\tilde\imath$ was estimated as the largest term in the sum times the bounded number of terms.

The second term in (3.12) can be bounded using Gershgorin's circle theorem and the facts that E is sparse and all of its entries are $\lesssim_{\varepsilon} \exp\big(-g_{\mathcal{E}}(z)\,c\big)$, which yields $\|E\|_2 \lesssim_{\varepsilon} \exp\big(-g_{\mathcal{E}}(z)\,c\big)$. Combining the bound on the first term of (3.12) from the previous paragraph with the above estimate on the second term, we obtain

$$\begin{aligned}
\big|\big((H-z)^{-1} - (H+E-z)^{-1}\big)(i,j)\big| &\lesssim_{\varepsilon} \exp\big(-2\,g_{\mathcal{E}}(z)\,c\big) + \exp\big(-2\,g_{\mathcal{E}}(z)\,c\big) \\
&= \exp\big(-2\,g_{\mathcal{E}}(z)\,c\big)
\end{aligned}$$

as claimed. □

3.3.2 Incomplete Selected Inversion

We propose the following incomplete algorithm for the selected inversion step.

Algorithm 3.5 Incomplete selected inversion

```
for $j = m, \dots, 1$ do
    $B(\tilde{r}, j) = -B(\tilde{r}, \tilde{r})\, \tilde{L}(\tilde{r}, j)$
    $B(j, \tilde{r}) = B(\tilde{r}, j)^T$
    $B(j, j) = \tilde{D}(j, j)^{-1} - B(j, \tilde{r})\, \tilde{L}(\tilde{r}, j)$
    $F(r \setminus \tilde{r}, j) = B(r \setminus \tilde{r}, \tilde{r})\, \tilde{L}(\tilde{r}, j)$
    $F(j, r \setminus \tilde{r}) = F(r \setminus \tilde{r}, j)^T$
    $F(j, j) = F(j, \tilde{r})\, \tilde{L}(\tilde{r}, j)$
end for
```

As in Algorithm 3.4, keeping track of the dropped F is not required in an actual implementation but doing so facilitates our discussion of the errors committed by this algorithm.

Theorem 3.3.11 *Algorithm* 3.5 *requires* $\mathcal{O}(m)$ *runtime and memory.*

Proof. Analogous to Theorem 3.3.3. □

The analysis of this algorithm proceeds along the same lines as in Subsection 3.3.1: we first establish an a-posteriori bound in terms of the dropped entries F in Theorem 3.3.13, then we argue that $|F(i,j)|$ should decay like $|A^{-1}(i,j)|$ in Conjecture 3.3.14, and finally we derive an a-priori bound based on this conjecture in Corollary 3.3.15. For all of these steps, we will need the following result which establishes that $\tilde{A}^{-1} = (A+E)^{-1}$ exhibits the same localisation as A^{-1}.

Lemma 3.3.12 *In the notation of Definition* 3.3.2 *and assuming Conjecture* 3.3.8*, we have that*

$$|\tilde{A}^{-1}(i,j)| \lesssim_\varepsilon \exp\bigl(-g_{\mathcal{E}}(z)\, d(i,j)\bigr).$$

Proof. According to Theorem 3.3.5 (a-posteriori error bound for $\tilde{A}^{-1}$), Theorem 3.3.6 (sparsity of E) and Conjecture 3.3.8 (localisation of E), we have that

$$|\tilde{A}^{-1}(i,j)| \lesssim_\varepsilon |A^{-1}(i,j)| + \sum_{(\tilde{\imath},\tilde{\jmath}) \in \mathrm{nz}(E)} \exp\Bigl(-g_{\mathcal{E}}(z)\, \bigl(d(i,\tilde{\imath}) + c + d(\tilde{\jmath}, j)\bigr)\Bigr).$$

The claim follows by estimating the first term on the right hand side using Theorem 3.2.2 (localisation of A^{-1}) and the second term using the boundedness of geometric series. □

Theorem 3.3.13 *In the notation of Definition* 3.3.2 *and assuming Conjecture* 3.3.8*, we have that*

$$\Big|\big(\tilde{A}^{-1} - B\big)(i,j)\Big| \lesssim_\varepsilon \sum_{\tilde{\jmath}=j+1}^{m} \exp\big(-g_{\mathcal{E}}(z)\,\mathrm{level}(\tilde{\jmath},j)\big)\,|F(i,\tilde{\jmath})| + \dots$$
$$+ \sum_{\tilde{\imath},\tilde{\jmath}=j+1}^{m} \exp\Big(-g_{\mathcal{E}}(z)\,\big(\mathrm{level}(i,\tilde{\imath}) + \mathrm{level}(\tilde{\jmath},j)\big)\Big)\,|F(\tilde{\imath},\tilde{\jmath})|.$$

This bound is illustrated in Figure 3.4b*.*

Proof. Let us first consider the application of Algorithm 3.3 (exact selected inversion) to the matrix $\tilde{A} = A + E = \tilde{L}\tilde{D}\tilde{L}^T$. We note that the entries $\tilde{A}^{-1}(i',j')$ computed by this algorithm after iteration j depend only on $\tilde{L}(\cdot,\ell)$, $\tilde{D}(\ell,\ell)$ and $\tilde{A}^{-1}(\bar{r},\bar{r})$, hence iterations $j' = j-1,\dots,1$ may be interpreted as a map $\phi : \big(\tilde{L}(\cdot,\ell), \tilde{D}(\ell,\ell), \tilde{A}^{-1}(\bar{r},\bar{r})\big) \mapsto \tilde{A}^{-1}$ which must be unique since the map from $\tilde{A}^{-1}$ to $\big(\tilde{L}(\cdot,\ell), \tilde{D}(\ell,\ell), \tilde{A}^{-1}(\bar{r},\bar{r})\big)$ is injective. This uniqueness allows us to determine ϕ by applying the selected inversion subalgorithm to the block-LDL^T factorisation from (3.3), which yields

$$\tilde{A}^{-1} = \begin{pmatrix} \tilde{A}(\ell,\ell)^{-1} + \tilde{A}(\ell,\ell)^{-1}\tilde{A}(\ell,\bar{r})\,\tilde{A}^{-1}(\bar{r},\bar{r})\,\tilde{A}(\bar{r},\ell)\,\tilde{A}(\ell,\ell)^{-1} & -\tilde{A}(\ell,\ell)^{-1}\tilde{A}(\ell,\bar{r})\,\tilde{A}^{-1}(\bar{r},\bar{r}) \\ -\tilde{A}^{-1}(\bar{r},\bar{r})\,\tilde{A}(\bar{r},\ell)\,\tilde{A}(\ell,\ell)^{-1} & \tilde{A}^{-1}(\bar{r},\bar{r}) \end{pmatrix}. \tag{3.13}$$

Note that this is indeed a map in terms of $\tilde{L}(\cdot,\ell)$, $\tilde{D}(\ell,\ell)$ since all of the submatrices in (3.13) other than $\tilde{A}(\bar{r},\bar{r})^{-1}$ can be computed from $\tilde{L}(\cdot,\ell)$, $\tilde{D}(\ell,\ell)$.

Let us now assume for the moment that Algorithm 3.5 (incomplete selected inversion) only drops entries in $B\big(\bar{r},j\big)$ and $B\big(j,\bar{r}\big)$ such that[2]

$$B(\bar{r},\bar{r}) = \tilde{A}^{-1}(\bar{r},\bar{r}) + F(\bar{r},\bar{r}).$$

Since by assumption the incomplete inversion does not perform any additional mistakes after iteration j, we have that $B = \phi\big(B(\bar{r},\bar{r})\big)$ where for brevity we dropped the arguments of ϕ other than $\tilde{A}^{-1}(\bar{r},\bar{r})$, and since ϕ is affine in $\tilde{A}^{-1}(\bar{r},\bar{r})$ it further follows that

$$\begin{aligned} B = \phi\big(B(\bar{r},\bar{r})\big) &= \phi\big(\tilde{A}^{-1}(\bar{r},\bar{r})\big) + \phi\big(F(\bar{r},\bar{r})\big) - \phi(0) \\ &= \qquad \tilde{A}^{-1} \qquad + \phi\big(F(\bar{r},\bar{r})\big) - \phi(0). \end{aligned} \tag{3.14}$$

In the simplified case where errors occur only in $B\big(\bar{r},j\big)$ and $B\big(j,\bar{r}\big)$, the claim

[2] We would like to emphasise that this simple formula only holds for the first iteration j where entries are dropped, since in later iterations $j' < j$ the error introduced by the dropped entries may propagate into other entries of B.

then follows by estimating the entries of $\phi\big(F(\bar r, \bar r)\big) - \phi(0)$ using the localisation of $\tilde A(\ell,\ell)^{-1}$ described in Lemma 3.3.12, and the general estimate follows by applying (3.14) recursively for each j. □

Conjecture 3.3.14 *In the notation of Definition* 3.3.2*, we have that*

$$|F(i,j)| \lesssim_\varepsilon \exp\big(-g_{\mathcal{E}}(z)\, d(i,j)\big).$$

Discussion. From the proof of Theorem 3.3.13, it follows that F can be computed recursively according to

$$F(i,j) = \tilde A^{-1}(i,j) - F\big(i,r(i)\big)\, M^T\big(r(i),j\big) + M\big(i,r(j)\big)\, F\big(r(j),r(j)\big)\, M^T\big(r(j),j\big)$$

where

$$M(i,\tilde\imath) = \begin{cases} A_{\ell(\tilde\imath)}^{-1}\big(i,\ell(\tilde\imath)\big)\, A\big(\ell(\tilde\imath),\tilde\imath\big) & i < \tilde\imath \\ 0 & \text{otherwise} \end{cases}$$

and $A_\ell := A(\ell,\ell)$. Proving Conjecture 3.3.14 thus faces the same obstacle as Conjecture 3.3.8, namely that the errors committed at iteration j depend on errors committed at previous iterations $j' > j$ such that any bound deteriorates exponentially in the recursion depth. □

Corollary 3.3.15 *In the notation of Definition* 3.3.2*, and assuming Conjecture* 3.3.8*, Assumption* 3.3.9 *and Conjecture* 3.3.14*, we have that*

$$\big\|(H+E-z)^{-1} - B^{-1}\big\|_{\mathrm{nz}(H)} \lesssim_\varepsilon \exp\big(-2\, g_{\mathcal{E}}(z)\, c\big).$$

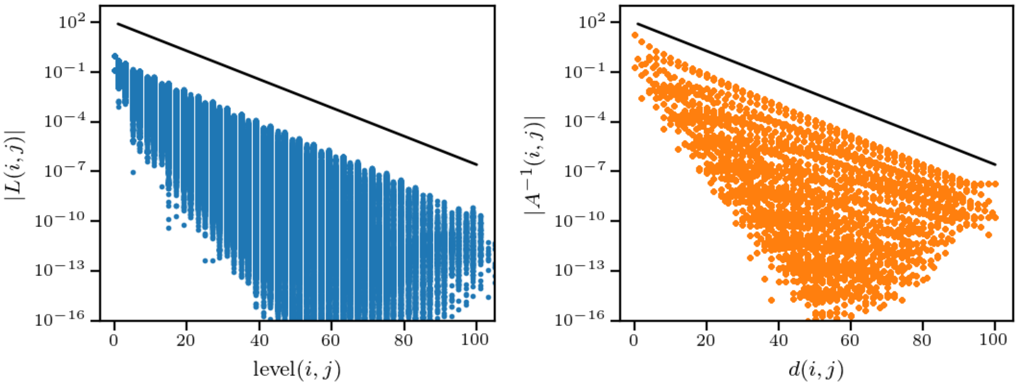
Proof. Analogous to Corollary 3.3.10. □

As noted in (3.8), the total error of the incomplete selected inversion algorithm is upper-bounded by the sum of the errors of the two substeps. Combining Theorems 3.3.3 and 3.3.11 and Corollaries 3.3.10 and 3.3.15 thus yields the following theorem which summarises the main result of this section.

Theorem 3.3.16 *In the notation of Definition* 3.3.2*, and assuming Conjecture* 3.3.8*, Assumption* 3.3.9 *and Conjecture* 3.3.14*, we have that*

$$\big\|(H-z)^{-1} - B\big\|_{\mathrm{nz}(H)} \lesssim_\varepsilon \exp\big(-2\, g_{\mathcal{E}}(z)\, c\big).$$

Furthermore, B can be computed in $\mathcal{O}(m)$ runtime and memory.



(a) Factorisation

(b) Inverse

Figure 3.5: Localisation of the L-factor (left) and inverse of the matrix $H - 0.98$ with H as in (3.15) and $d = 2$. The black lines indicate the rate of decay $g_{\mathcal{E}}(z)$ predicted by Theorems 3.2.2 and 3.2.5.

3.4 Numerical Experiments

This section illustrates the theory presented above at the example of a toy Hamiltonian $H \in \mathbb{R}^{m \times m}$ with entries $H(i,j)$ given by

$$H(i,j) := \begin{cases} (-1)^{d(i,1)} & \text{if } i = j, \\ -\frac{1}{2d} & \text{if } i \sim j, \\ 0 & \text{otherwise,} \end{cases} \tag{3.15}$$

where $d \in \{1,2,3\}$ denotes the dimension and $i \sim j$ if i and j are nearest neighbours in a d-dimensional Cartesian mesh with periodic boundary conditions. We note that the off-diagonal entries $H(i,j) = -\frac{1}{2d}$ correspond to a shifted and scaled finite-difference discretisation of the d-dimensional Laplace operator, and the diagonal entries $H(i,i) = (-1)^{d(i,1)}$ take the form of a chequerboard pattern where each vertex has the opposite sign compared to its neighbours. In two and three dimensions, we use an approximate minimum degree (AMD) ordering [ADD96, ADD04] to improve the sparsity of the LDL^T factorisation, while in one dimension we use a simple left-to-right ordering as shown in Figure 3.1. Details regarding the hardware and software used for these experiments are provided in Appendix A.1.

Example 3.4.1 (localisation) The chequerboard pattern along the diagonal causes the spectrum $\mathcal{E}$ of H to split into two intervals, $\mathcal{E} := [-\sqrt{2}, -1] \cup [1, \sqrt{2}]$, such that one may think of H as the Hamiltonian matrix of an insulator with band gap

$(-1,1)$. Figure 3.5 compares the entries of the LDL^T factorisation and inverse of $H - z$ against the predictions of Theorems 3.2.2 and 3.2.5 for a shift z in the band gap, and we observe that the theory is matched perfectly. In particular, the entries of the L-factor decay with the same rate $g_{\mathcal{E}}(z)$ as the inverse, which indicates that the spectra of all leading submatrices of H are indeed contained in $\mathcal{E}$ as conjectured in Remark 3.2.6.

The excellent agreement between the theoretical and empirical convergence rates is a consequence of the simple sparsity pattern in (3.15), and the agreement may be worse for a more realistic Hamiltonian.

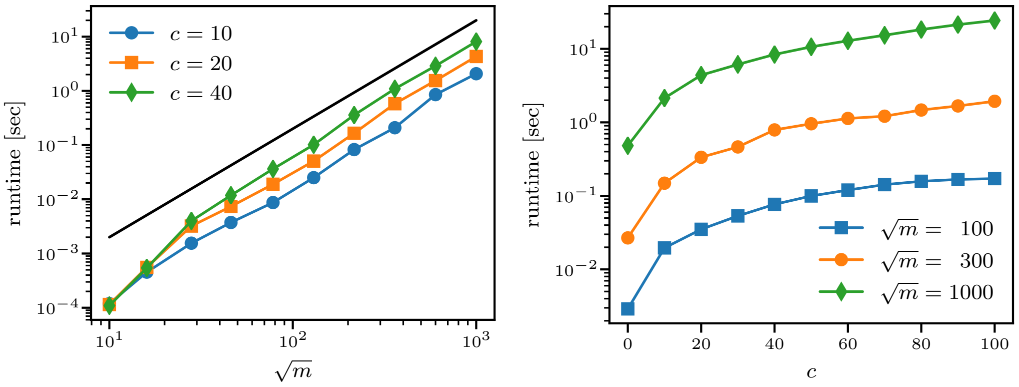
Example 3.4.2 (convergence) We recall from Theorem 3.3.16 that the incomplete selected inversion algorithm is predicted to scale linearly in m and converge exponentially in the cut-off level-of-fill c with a rate of convergence equal to twice the localisation rate $g_{\mathcal{E}}(z)$. These theoretical findings are confirmed numerically in Figures 3.6 and 3.7, respectively. The staircase pattern in Figure 3.7b is a consequence of the AMD ordering arranging the vertices such that all $\mathrm{level}(i,j)$ are either 0, ∞ or odd which yields $\mathrm{ifnz}_{c=2k}(H) = \mathrm{ifnz}_{c=2k+1}(H)$ for all $k \in \mathbb{N}_{>0}$.

Example 3.4.3 Finally, we continue the discussion of Assumption 3.3.9 ($d(i,j) \sim \mathrm{level}(i,j)$) which was used to derive Theorem 3.3.16 but which we have seen to be violated in the case of one-dimensional periodic chains. Figure 3.8 shows that even in this case, the incomplete algorithms converge with rate $2\,g_{\mathcal{E}}(z)$ as predicted by Theorem 3.3.16, but the convergence breaks down at a cut-off level-of-fill c of about $\frac{m}{2}$ after which the error stagnates.

In the framework of Section 3.3, this observation may be explained as follows. Due to the simple graph structure of H, the matrix of dropped entries E from Algorithm 3.4 contains precisely two nonzero entries at locations $i = m$, $j = c + 2$ and the transpose thereof, and by Conjecture 3.3.8 these entries satisfy

$$|E(m, c+2)| \lesssim_{\varepsilon} \exp\big(-g_{\mathcal{E}}(z)\,\mathrm{level}(m, c+2)\big) = \exp\big(-g_{\mathcal{E}}(z)\,(c+1)\big).$$

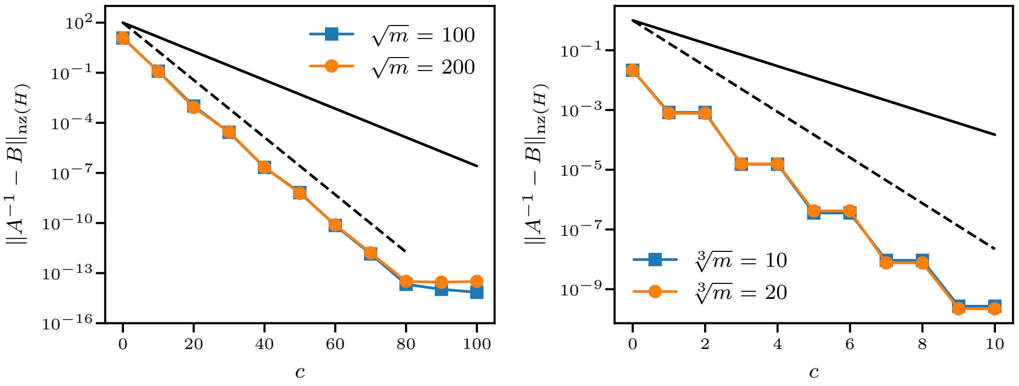
(In this case, Conjecture 3.3.8 can easily be proven since the incomplete factorisation algorithm only drops a single entry.) According to Theorem 3.3.5, the error due to



(a) Scaling in m

(b) Scaling in c

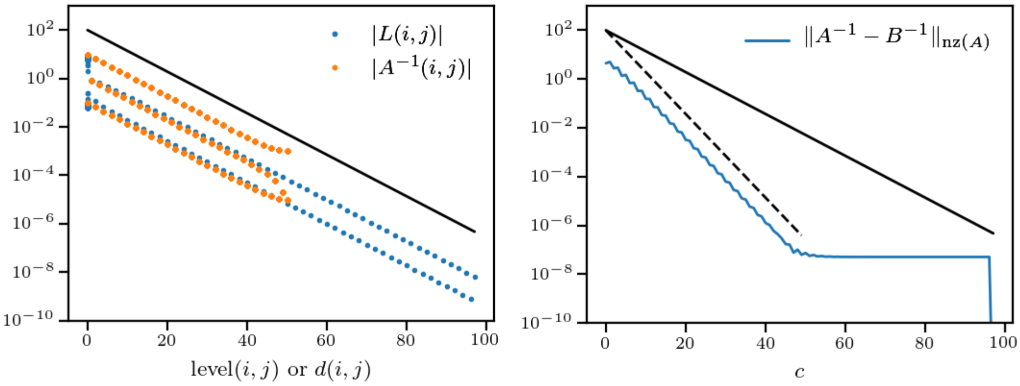
Figure 3.6: Scaling of the incomplete selected inversion algorithm applied to the matrix H from (3.15) with respect to the matrix size m (left) and cut-off level-of-fill c (right). The black line in (a) indicates $\mathcal{O}(m)$ scaling. The reported runtimes are the minimum out of three runs for each pair m, c.



(a) $d = 2, z = 0.98$

(b) $d = 3, z = 0$

Figure 3.7: Error vs. cut-off level-of-fill c of incomplete factorisation and selected inversion algorithm applied to $A = H - z$ with H as in (3.15). The solid black lines indicate exponential decay with rate $g_{\mathcal{E}}(z)$, and the dashed lines indicate twice this rate.



(a) Localisation

(b) Error

Figure 3.8: Localisation (left) and convergence of incomplete factorisation and selected inversion (right) of the matrix $H - z$ with H as in (3.15), $d = 1$ $m = 100$ and $z = 0.98$. The solid black lines indicate exponential decay with rate $g_{\mathcal{E}}(z)$, and the dashed line indicates twice this rate.

the incomplete factorisation is thus upper-bounded by

$$\begin{aligned} \|A^{-1} - \tilde{A}^{-1}\|_{\mathrm{nz}(A)} &\lesssim_{\varepsilon} \max_{(i,j)\in \mathrm{nz}(A)} \exp\Big(-g_{\mathcal{E}}(z)\, \big(d(i,m) + c + 1 + d(c+2,j)\big)\Big) \\ &\lesssim_{\varepsilon} \exp\Big(-g_{\mathcal{E}}(z)\, \big(c + 1 + \min\{d(c+2,1), d(c+2,m-1)\}\big)\Big) \\ &= \begin{cases} \exp\big(-2\, g_{\mathcal{E}}(z)\, (c+1)\big) & \text{if } c \le \frac{m-4}{2}, \\ \exp\big(-g_{\mathcal{E}}(z)\, (m-2)\big) & \text{otherwise,} \end{cases} \end{aligned} \tag{3.16}$$

and a similar bound can be derived for the error $\|\tilde{A}^{-1} - B\|_{\mathrm{nz}(A)}$ introduced by the incomplete selected inversion step. We note that (3.16) describes precisely the behaviour observed in Figure 3.8b.

As discussed after Assumption 3.3.9, we expect that $d(i,j) \sim \mathrm{level}(i,j)$ is rarely violated in dimensions $d > 2$ and if a reasonable (e.g. AMD or nested dissection) vertex ordering is used. This example further demonstrates that even if Assumption 3.3.9 is violated, the incomplete selected inversion algorithm still converges at the rate $2g_{\mathcal{E}}(z)$ until the cut-off level-of-fill c becomes $\mathcal{O}(m)$, at which point the speedup of the incomplete selected inversion compared to the exact algorithm vanishes anyway.

3.5 Conclusion

We have shown that the LDL^T factorisation of a sparse, well-conditioned[3] matrix A exhibits a localisation property similar to that of A^{-1}, and we have developed algorithms which exploit this property to compute selected entries of A^{-1} in $\mathcal{O}(m)$ runtime and memory. This opens up a new class of linear-scaling electronic structure algorithms based on rational approximation which we expect to be highly competitive compared to polynomial algorithms and optimisation algorithms based on the conjugate gradient method due to the following reasons.

- Like polynomial algorithms, the conjugate gradient iteration applied to the functional (1.15) uses only matrix products and sums and its convergence rate depends algebraically on the inverse temperature β and the band gap δE [Goe99]. We therefore expect that the following remarks comparing the polynomial and rational algorithms also apply to the comparison between rational and optimisation algorithms.

- We have seen in Chapter 2 that the rational degree required to approximate the Fermi-Dirac function to a fixed accuracy scales better than logarithmically in the temperature and band gap, compared to a linear dependence in the case of polynomial approximation. For low temperatures and small band gaps, the rational degree will therefore be orders of magnitude smaller than the polynomial degree.

- The cost of evaluating a rational approximation scales linearly in the number of poles, while the cost of evaluating a matrix polynomial scales quadratically in the degree due to the reduced sparsity in higher powers of H. Combined with the fast inversion algorithm developed in this chapter, this implies that rational functions may well be cheaper to evaluate than polynomials of a comparable degree. As a case in point, we mention that the selected inversion algorithm applied to the matrix (3.15) with $d = 2$, $\sqrt{m} = 300$ and $c = 20$ takes 0.4 seconds while evaluating the 20th power of the same matrix H takes 9 seconds (see Appendix A.1 for details regarding hardware and software). Evaluating a power of H with similar localisation properties as H^{-1} is thus over 20 times slower for these particular parameters, and this ratio will tip even further in favour of the selected inversion algorithm as we increase the localisation length c.

[3] We call a matrix A well-conditioned if its "smoothed" spectrum $\mathcal{E}$ (cf. Section 1.3) does not contain the origin.

What is needed next in order to realise the promised advantages of rational electronic structure algorithms is a massively parallel high-performance implementation of the incomplete selected inversion algorithm comparable to that presented for the exact algorithm in [JLY16]. Developing such a code will be the topic of future work, but we would like to point out that the parallelisation strategies from [JLY16] also apply to the incomplete factorisation and selected inversion algorithms and hence we expect similar parallel scaling.

Closely related work regarding the parallel implementation of the incomplete LU factorisation with arbitrary level-of-fills (as opposed to the more wide-spread ILU(0) and ILU(1) factorisations) can be found in [KK97, HP01, SZW03, DC11]. Furthermore, an alternative ILU algorithm based on iterative refinement of a trial factorisation and designed specifically to improve parallel scaling has recently been proposed in [CP15]. While this algorithm was found to be highly effective at finding factorisations suitable for preconditioning, it is unclear whether it is applicable in the context of the selected inversion algorithm where the accuracy requirements are much more stringent. Additionally, the algorithm from [CP15] will only lead to an asymptotic speedup for the selected inversion algorithm if a similar highly parallelisable algorithm for the selected inversion step can be developed, and it is not obvious whether such an algorithm exists since the algorithm from [CP15] is based on Theorem 3.3.4 which has no analogue in the selected inversion step.

Chapter 4

Approximation of the Conductivity Function

We recall from Section 1.2 that the conductivity tensor

$$\sigma = \sigma_{a,b} = \sum_{i_1,i_2} F_\zeta(\varepsilon_{i_1}, \varepsilon_{i_2})\, \langle \psi_{i_1}|M_a|\psi_{i_2}\rangle\, \langle \psi_{i_2}|M_b|\psi_{i_1}\rangle, \qquad a,b \in \{1,2,3\}, \tag{4.1}$$

with

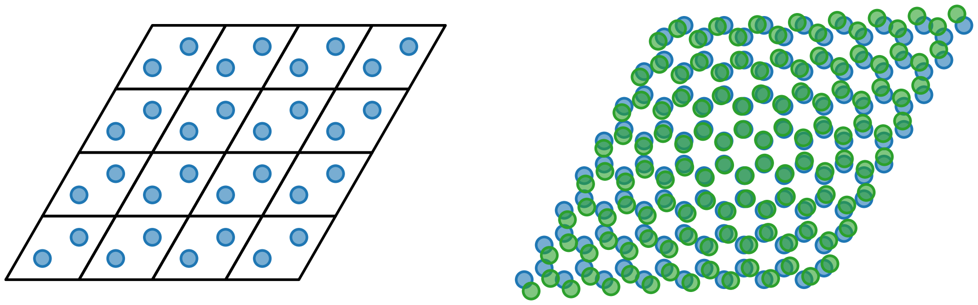
$$F_\zeta(E_1,E_2) = \frac{f_{\beta,E_F}(E_1) - f_{\beta,E_F}(E_2)}{E_1 - E_2}\, \frac{1}{E_1 - E_2 + \omega + i\eta} \tag{4.2}$$

expresses the linear relationship $\vec{J} = \sigma\, \vec{E}$ between the electric field $\vec{E}$ and the induced current $\vec{J}$. This quantity depends on the four parameters $\zeta = (\beta, E_F, \omega, \eta)$ with $\beta > 0$, $E_F \in (-1,1)$, $\omega, \eta > 0$ and the matrices $H, M_a \in \mathbb{C}^{m\times m}$ whose physical interpretations are described in Chapter 1. The variables ε_i, ψ_i in (4.1) denote the eigenvalues and -vectors, respectively, of the Hamiltonian matrix $H \in \mathbb{R}^{m\times m}$ which is assumed to be shifted and scaled such that its spectrum $\{\varepsilon_i\}$ is contained in $\mathcal{E} = [-1,1]$, cf. Section 1.3.

Conductivity is most commonly studied for crystalline materials where the atoms are arranged according to the blueprint of a unit cell which repeats itself infinitely often in all directions, see Figure 4.1a. We model such a system through two infinite vectors $y_{I,\alpha}$, $Z_{I,\alpha}$ with lattice index $I \in \mathbb{Z}^3$ and local index $\alpha \in \{1,\dots,N^{(0)}\}$, where as in Section 1.1 $y_{I,\alpha}$ represents the atomic coordinates and $Z_{I,\alpha}$ represents the atomic charges. These vectors are of the special form

$$y_{I,\alpha} = A\,I + y_{0,\alpha} \qquad \text{and} \qquad Z_{I,\alpha} = Z_{0,\alpha},$$

with $A \in \mathbb{R}^{3\times 3}$ an invertible matrix specifying the shifts between adjacent unit cells



(a) Monolayer

(b) Twisted bilayer

Figure 4.1: Arrangement of atoms (coloured dots) in (a) a graphene monolayer and (b) a graphene bilayer where the two sheets are rotated by three degrees relative to each other. The black parallelograms in (a) indicate the unit cells.

and $y_{0,\alpha} \in A\,[0,1]^3$, $Z_{0,\alpha} \in \mathbb{R}$ defining the arrangement of the atoms within a unit cell. As before, we assume the Hamiltonian of such a system to be given as a sparse matrix, but due to the infinite number of atoms, this matrix has now an infinite number of entries $H_{I,\alpha;J,\beta}$ with $I, J \in \mathbb{Z}^3$ and $\alpha, \beta \in \{1, \ldots, m^{(0)}\}$ and sums over the eigenpairs of H have to be interpreted as integrals over the spectral measure of H. Under these circumstances, the diagonalisation algorithm from Section 1.3 becomes highly competitive since Bloch's theorem (an extension of the theory of Fourier series) allows us to sample from the spectral measure of the infinite Hamiltonian H by only solving an eigenproblem in terms of a finite unit-cell Hamiltonian $H^{(0)}_{\alpha,\beta}$. However, Bloch's theorem breaks down as soon as the periodicity is lost, and new algorithms must be developed to fill the gap.

This chapter presents such an algorithm for computing the conductivity of incommensurate bilayers, which are stackings of two atomically thin sheets such that their combination is aperiodic even though each individual sheet is periodic. The aperiodicity arises, for example, due to mismatching lattice constants (e.g. two one-dimensional lattices with unit cells of length 1 and π, respectively), or because one of the layers is rotated with respect to the other as illustrated in Figure 4.1b. Multilayer systems in general and incommensurate bilayers in particular have received much attention in the physics literature in recent years since they promise to provide a Lego-like toolbox for designing new materials with highly unusual properties [GG13]. Most notably, it has been observed in [CFF$^+$18] that twisted graphene bilayers as shown in Figure 4.1b exhibit superconductivity for certain "magic" twist angles. We hope that the mathematical tools presented here and in the closely related publication [MLO17] can help to explain such phenomena and guide further exploration in this field.

It has been observed in [MLO17] that the mismatch between the two sheets introduces a form of ergodicity which allows us to reduce computations on the infinite bilayer system to equivalent computations on just the two unit cells in a way similar to Bloch's theorem, and our algorithm will be based on the application of this result to the conductivity which we will discuss in Section 4.1. Unlike Bloch's theorem, however, the unit-cell computations in the case of incommensurate bilayers require padding with a buffer region, which often leads to system sizes far beyond the reach of the diagonalisation algorithm. In Section 4.2, we will propose an alternative, linear-scaling algorithm for the unit-cell problem based on the function approximation idea from Section 1.3. This algorithm will require tensor-product polynomial or rational approximations to the conductivity function $F_\zeta(E_1, E_2)$ from (4.2), which introduces a new challenge to the theory presented in Chapter 2 since $F_\zeta(E_1, E_2)$ is a two-dimensional function. Our main contribution in this chapter is the detailed analysis of this approximation problem, which will be carried out in Section 4.4 for the polynomial case and in Section 4.5 for the rational case. Finally, we will remark on the practical implementation of the proposed scheme in Section 4.6.

Disclaimer. The content of this chapter is the result of a collaboration between Mitchell Luskin at the University of Minnesota, Daniel Massatt at the University of Chicago, Christoph Ortner and myself. The decomposition into local conductivities presented in Section 4.1 has been developed by my collaborators and is presented here solely for the purpose of motivating the approximation problem discussed in the remaining sections. The material presented in Sections 4.2, 4.4 and 4.5 is my own work but received useful inputs from the aforementioned collaborators. This write-up has been authored by myself, and parts of it will appear in our forthcoming publication [EMLO19].

4.1 Incommensurate Bilayers and Local Conductivity

Ignoring the component in the third direction, the vectors of atomic coordinates and charges in a bilayer system are given by

$$y_{\ell,I,\alpha_\ell} := A_\ell \, I + y^{(0)}_{\ell,\alpha_\ell} \qquad \text{and} \qquad Z_{\ell,I,\alpha_\ell} := Z^{(0)}_{\ell,\alpha_\ell},$$

respectively, where $\ell \in \{1,2\}$ enumerates the sheets and $I \in \mathbb{Z}^2$, $\alpha_\ell \in \{1,\dots,N_\ell^{(0)}\}$, $A_\ell \in \mathbb{R}^{2\times 2}$, $y_{\ell,\alpha_\ell}^{(0)} \in A_\ell\,[0,1]^2$, $Z_{\ell,\alpha_\ell}^{(0)} \in \mathbb{R}$ as above. We call such a system incommensurate if the following condition is satisfied.

Definition 4.1.1 *Two matrices $A_1, A_2 \in \mathbb{R}^2$ are called* incommensurate *if*

$$A_1^{-T}\,\mathbb{Z}^2 \cap A_2^{-T}\,\mathbb{Z}^2 = \{0\}.$$

Remark 4.1.2 A_1, A_2 being incommensurate according to Definition 4.1.1 does not imply $A_1\,\mathbb{Z}^2 \cap A_2\,\mathbb{Z}^2 = \{0\}$. To see this, let us consider the counterexample

$$A_1 = \mathbb{I}, \qquad A_2 = \begin{pmatrix} 1+\sqrt{2} & \sqrt{2} \\ \sqrt{3} & \sqrt{3} \end{pmatrix}$$

for which we have that

$$A_1\,\mathbb{Z}^2 \cap A_2\,\mathbb{Z}^2 = \left\{ \begin{pmatrix} k \\ 0 \end{pmatrix} = A_2 \begin{pmatrix} k \\ -k \end{pmatrix} \;\middle|\; k \in \mathbb{Z} \right\}$$

but

$$A_1^{-T}\,\mathbb{Z}^2 \cap A_2^{-T}\,\mathbb{Z}^2 = A_2^{-T}\left(A_2^T\,\mathbb{Z}^2 \cap \mathbb{Z}^2\right) = \{0\}$$

since the second entry of $A_2^T\,k$ equals $\sqrt{2}\,k_1 + \sqrt{3}\,k_2 \notin \mathbb{Z}$ for all $k \in \mathbb{Z}^2 \setminus \{0\}$. We conclude that strictly speaking, aperiodicity does not imply incommensurability or vice versa, but we expect that the difference is of little relevance in practice.

This example has been taken from [TRM].

Incommensurability will be important for our purposes due to the following result.

Theorem 4.1.3 ([MLO17, Theorem 2.1]) *Let A_1, A_2 be incommensurate and let $g \in C_{\mathrm{per}}\big(A_2\,[0,1]^2\big)$. We then have that*

$$\overline{\sum_{I\in\mathbb{Z}^2}} g\big(A_1\,I\big) = \frac{1}{\det(A_2)} \int_{[0,1]^2} g\big(A_2\,u\big)\,du, \tag{4.3}$$

where

$$\overline{\sum_{I\in\mathbb{Z}^2}} q_I := \lim_{n\to\infty} \frac{1}{(2n+1)^2} \sum_{I\in\{-n,\dots,n\}^2} q_I.$$

Theorem 4.1.3 suggests that we evaluate a single entry $\sigma = \sigma_{a,b}$ (we drop the subscripts a, b in the following for brevity) of the conductivity tensor from (4.1) according to the following outline, which mirrors the propositions in [MLO17] for the density of states of incommensurate bilayers and is made rigorous in [EMLO19].

1. Introduce a local conductivity

$$\sigma_{\ell,I,\alpha_\ell} := \sum_{i_1,i_2} F_\zeta(\varepsilon_{i_1},\varepsilon_{i_2})\, \langle\psi_{i_1}|M_a|\psi_{i_2}\rangle\, \langle\psi_{i_2}|M_b|e_{\ell,I,\alpha_\ell}\rangle\langle e_{\ell,I,\alpha_\ell}|\psi_{i_1}\rangle \tag{4.4}$$

   such that

$$\sigma = \sum_{\ell=1}^{2} \overline{\sum_{I\in\mathbb{Z}^2}} \sum_{\alpha_\ell=1}^{m_\ell^{(0)}} \sigma_{\ell,I,\alpha_\ell}. \tag{4.5}$$

   Equation (4.5) follows from the identity $\mathbb{I} = \sum_{\ell,I,\alpha} |e_{\ell,I,\alpha}\rangle\langle e_{\ell,I,\alpha}|$.

2. Note that due to translation invariance, any physical quantity $q_{1,I}$ associated with unit cell I of sheet 1 can only depend on the shift $A_2\, u = \operatorname{mod}_{A_2}\big(A_1\, I\big)$ of said unit cell with respect to the unit cell of sheet 2 and vice versa, where $\operatorname{mod}_A(x) := x - A\, I$ with $I \in \mathbb{Z}^2$ such that $\operatorname{mod}_A(x) \in A\,[0,1]^2$. In particular, there must exist a function $\sigma_{1,\alpha_1}(A_2\, u)$ periodic on $A_2\,[0,1]^2$ such that the local conductivity σ_{1,I,α_1} can be written as

$$\sigma_{1,I,\alpha_1} = \sigma_{1,\alpha_1}\big(\operatorname{mod}_{A_2}(A_1\, I)\big),$$

   and likewise for σ_{2,I,α_2}.

3. Use Theorem 4.1.3 to rewrite the sum (4.5) as

$$\sigma = \sum_{\ell=1}^{2} \frac{1}{\det\big(A_{T(\ell)}\big)} \int_{u\in[0,1]^2} \sum_{\alpha_\ell=1}^{m_\ell^{(0)}} \sigma_{\ell,\alpha_\ell}\big(A_{T(\ell)}\, u\big)\, du$$

   where $T(\ell)$ denotes the transposition $T(1) = 2$ and $T(2) = 1$.

4. Approximate the integrals in (4.5) using some quadrature rule $(u_i, w_i)_{i=1}^q$ such that

$$\sum_{i=1}^{q} w_i\, \sigma_{\ell,\alpha_\ell}\big(A_{T(\ell)}\, u_i\big) \approx \int_{u\in[0,1]^2} \sigma_{\ell,\alpha_\ell}\big(A_{T(\ell)}\, u\big)$$

   A detailed analysis of this approximation will be the topic of future work, but we expect that exponential convergence is achievable and we will assume this convergence in the following for ease of exposition.

The last step reduces the problem of computing the overall conductivity σ to that of sampling local conductivities $\sigma_{\ell,\alpha_\ell}(A_{T(\ell)}\, u)$ at some finite number of quadrature points u. Following the domain decomposition approach from Section 1.3, we propose to compute these samples $\sigma_{1,\alpha_1}(A_2\, u)$ according to the following outline.

1. Assemble a local configuration

$$y^r_{1,I_1,\alpha_1} := y_{1,I_1,\alpha_1}, \qquad y^r_{2,I_2,\alpha_2} := y_{2,I_2,\alpha_2} - A_2\, u, \qquad Z^r_{\ell,I_\ell,\alpha_\ell} := Z_{\ell,I_\ell,\alpha_\ell}$$

   with $I_\ell \in \mathbb{Z}$ such that $|A_\ell I_\ell| \le r$ for some truncation radius $r > 0$, and $\alpha_\ell \in \{1, \dots, N^{(0)}_\ell\}$. This local configuration $y^r_{\ell,I_\ell,\alpha_\ell}$ corresponds to a circular cut-out of radius r around the origin, with the second lattice shifted by $-A_2\, u$ such that the origin of lattice 1 sees the same environment as the unit cell of sheet 1 with index I in the unshifted bilayer, where I and u are related by $A_2\, u = \operatorname{mod}_{A_2}(A_1\, I)$.

2. Assemble a local Hamiltonian $H^r_1(A_2\, u)$ and velocity matrices $M^r_{1,a}(A_2\, u)$ based on the above local configuration, and evaluate an approximate local conductivity $\sigma^r_{1,\alpha_1}\big(A_2 u\big) = \sigma^r_{1,0,\alpha_1}$ by inserting these matrices into the formula (4.4).

3. Proceed likewise for $\sigma^r_{2,\alpha_2}\big(A_1 u\big)$.

The approximate local conductivities computed in this way converge exponentially in the buffer radius r.

Theorem 4.1.4 *There exist $C, \gamma > 0$ depending only on ζ such that*

$$\big|\sigma^r_{\ell,\alpha_\ell}\big(A_{T(\ell)}\, u\big) - \sigma_{\ell,\alpha_\ell}\big(A_{T(\ell)}\, u\big)\big| \le C\, \exp(-\gamma\, r).$$

Proof. See Subsection 4.8.3. □

We have thus constructed an approximate conductivity

$$\sigma^{r,q} := \sum_{\ell \in \{1,2\}} \frac{1}{\det(A_{T(\ell)})} \sum_{i=1}^{q} w_i \sum_{\alpha=1}^{m^{(0)}_\ell} \sigma^r_{\ell,\alpha_\ell}\big(A_{T(\ell)}\, u_i\big) \approx \sigma$$

which is computable in finite time and converges exponentially in both the truncation radius r and the number of quadrature points q. However, the convergence rate with respect to the radius r deteriorates quickly for growing inverse temperatures β and shrinking inverse relaxation times η, such that in many applications the truncated configurations $y^r_{\ell,I_\ell,\alpha_\ell}$ must include tens of thousands of atoms in order to achieve acceptable accuracy and the diagonalisation algorithm becomes prohibitively expensive. In the next section, we will construct an alternative algorithm based on the function approximation idea from Section 1.3 which scales only linearly in the system size and is therefore a promising candidate for overcoming the scaling problem of the diagonalisation algorithm. To facilitate its presentation, we simplify the notation by writing σ_{loc}, H_{loc} and M_a^{loc} instead of $\sigma^r_{\ell,\alpha_\ell}\big(A_{T(\ell)}\,u_i\big)$, $H^r_\ell\big(A_{T(\ell)}\,u_i\big)$ and $M^r_{\ell,a}\big(A_{T(\ell)}\,u_i\big)$, respectively.

4.2 Local Conductivities via Chebyshev Approximation

Let us denote by $\tilde F_\zeta$ an approximate conductivity function obtained by truncating the Chebyshev series of the exact function F_ζ from (4.2), i.e.

$$\begin{aligned} \tilde F_\zeta(E_1, E_2) &:= \sum_{(k_1,k_2)\in K} c_{k_1k_2}\, T_{k_1}(E_1)\, T_{k_2}(E_2) \\ &\approx \sum_{k_1,k_2=0}^{\infty} c_{k_1k_2}\, T_{k_1}(E_1)\, T_{k_2}(E_2) = F_\zeta(E_1, E_2) \end{aligned} \tag{4.6}$$

where $K \subset \mathbb{N}^2$ is a finite set of indices and $T_k(E)$ denotes the kth Chebyshev polynomial defined through the three-term recurrence relation

$$T_0(x) = 1, \quad T_1(x) = x, \quad T_{k+1}(x) = 2x\, T_k(x) - T_{k-1}(x). \tag{4.7}$$

Replacing F_ζ with $\tilde F_\zeta$ in the local conductivity formula (4.4), we obtain an approximate local conductivity

$$\begin{aligned} \tilde\sigma_{\mathrm{loc}} &:= \sum_{i_1,i_2} \tilde F_\zeta(\varepsilon_{i_1}, \varepsilon_{i_2})\, \langle \psi_{i_1}|M_a^{\mathrm{loc}}|\psi_{i_2}\rangle\, \langle \psi_{i_2}|M_b^{\mathrm{loc}}|e_{\ell,0,\alpha_\ell}\rangle \langle e_{\ell,0,\alpha_\ell}|\psi_{i_1}\rangle \\ &= \sum_{i_1,i_2} \sum_{(k_1,k_2)\in K} c_{k_1k_2}\, \langle e_{\ell,I,\alpha_\ell}|\psi_{i_1}\rangle\, T_{k_1}(\varepsilon_{i_1})\, \langle \psi_{i_1}|M_a^{\mathrm{loc}}|\psi_{i_2}\rangle\, T_{k_2}(\varepsilon_{i_2})\, \langle \psi_{i_2}|M_b^{\mathrm{loc}}|e_{\ell,I,\alpha_\ell}\rangle \\ &= \sum_{(k_1,k_2)\in K} c_{k_1k_2} \Big(T_{k_1}(H_{\mathrm{loc}})\, M_a^{\mathrm{loc}}\, T_{k_2}(H_{\mathrm{loc}})\, M_b^{\mathrm{loc}} \Big)_{0,\alpha_\ell;0,\alpha_\ell} \end{aligned} \tag{4.8}$$

which can be evaluated without computing the eigendecomposition as follows.

Algorithm 4.1 Local conductivity via Chebyshev approximation

```
1: $|v_{k_1}\rangle := M_a^{\mathrm{loc}}\, T_{k_1}(H_{\mathrm{loc}})\, |e_{\ell,0,\alpha_\ell}\rangle$ for all $k_1 \in K_1 := \{k_1 \mid \exists k_2 : (k_1,k_2) \in K\}$.
2: $|w_{k_2}\rangle := T_{k_2}(H_{\mathrm{loc}})\, M_b^{\mathrm{loc}}\, |e_{\ell,0,\alpha_\ell}\rangle$ for all $k_2 \in K_2 := \{k_2 \mid \exists k_1 : (k_1,k_2) \in K\}$.
3: $\tilde\sigma_{\mathrm{loc}} := \sum_{(k_1,k_2)\in K} c_{k_1k_2}\, \langle v_{k_1}|w_{k_2}\rangle$.
```

Lines 1 and 2 of Algorithm 4.1 require $|K_1|$ and $|K_2|$, respectively, matrix-vector products when evaluated using the recurrence relation (4.7), while line 3 requires $|K|$ inner products. Due to the sparsity of $H_{\mathrm{loc}} \in \mathbb{R}^{m\times m}$, both types of products require

$\mathcal{O}(m)$ floating-point operations; thus we conclude that Algorithm 4.1 scales linearly in the matrix size m. Furthermore, the error in the computed local conductivity $\tilde\sigma_{\mathrm{loc}}$ can be estimated in terms of the dropped Chebyshev coefficients $c_{k_1k_2}$ as follows.

Lemma 4.2.1 *We have that*

$$\big|\tilde\sigma_{\mathrm{loc}} - \sigma_{\mathrm{loc}}\big| \lesssim \sum_{(k_1,k_2)\in\mathbb{N}^2\setminus K} |c_{k_1k_2}|.$$

Proof. The bound follows immediately from (4.8) after noting that M_a^{loc} and $T_k(H_{\mathrm{loc}})$ are bounded for all $a \in \{1,2\}$ and $k \in \mathbb{N}$. □

A more careful analysis of Algorithm 4.1 reveals that since $|K_1|, |K_2| \leq |K|$ and both matrix-vector and inner products require $\mathcal{O}(m)$ floating-point operations, the computational cost of this algorithm is dominated by the cost of line 3 which is $\mathcal{O}(|K|)$ inner products. The runtime of Algorithm 4.1 is thus minimized by choosing $|K|$ as small as possible subject to the constraint that $\sum_{(k_1,k_2)\in\mathbb{N}^2\setminus K} |c_{k_1k_2}|$ must be less than some error tolerance. The optimal choice for K is then to truncate the infinite Chebyshev series using some tolerance $\tau > 0$,

$$K(\tau) := \big\{(k_1,k_2) \in \mathbb{N}^2 \mid |c_{k_1k_2}| \geq \tau\big\};$$

thus the size of K is linked to the decay of the Chebyshev coefficients $c_{k_1k_2}$ which in turn depends on the regularity properties of F_ζ. To analyze these, it is convenient to split the conductivity function $F_\zeta(E_1,E_2) = f_{\mathrm{temp}}(E_1,E_2)\, f_{\mathrm{relax}}(E_1,E_2)$ from (4.2) into the two factors

$$f_{\mathrm{temp}}(E_1,E_2) := \frac{f_{\beta,E_F}(E_1) - f_{\beta,E_F}(E_2)}{E_1 - E_2} \tag{4.9}$$

and

$$f_{\mathrm{relax}}(E_1,E_2) := \frac{1}{E_1 - E_2 + \omega + i\eta} \tag{4.10}$$

where the subscripts reflect the fact that f_{temp} mainly depends on the inverse temperature β while f_{relax} mainly depends on the inverse relaxation time η (the Fermi energy E_F and frequency ω play only a minor role in the developments to come). These functions are easily seen to be analytic everywhere except, respectively, on the sets

$$\mathcal{S}_{\mathrm{temp}} := \big(\mathcal{S}_{\beta,E_F} \times \mathbb{C}\big) \cup \Big(\mathbb{C} \times \mathcal{S}_{\beta,E_F}\Big) \quad \text{with} \quad \mathcal{S}_{\beta,E_F} := \big\{E_F + \tfrac{i\pi k}{\beta} \mid k \text{ odd}\big\} \tag{4.11}$$

Constraint	Parameter range	# significant terms
Relaxation	$\beta \lesssim \eta^{-1/2}$	$\mathcal{O}\big(\eta^{-3/2}\big)$
Mixed	$\eta^{-1/2} \lesssim \beta \lesssim \eta^{-1}$	$\mathcal{O}\big(\beta\eta^{-1}\big)$
Temperature	$\eta^{-1} \lesssim \beta$	$\mathcal{O}\big(\beta^2\big)$

Table 4.1: Classification of conductivity parameters ζ and number of significant terms in the Chebyshev series of F_ζ.

(cf. (2.1)) and

$$\mathcal{S}_{\mathrm{relax}} := \big\{(E_2, E_2) \in \mathbb{C}^2 \mid E_1 - E_2 + \omega + i\eta = 0\big\};$$

hence the conductivity function F_ζ is analytic except on the union of these two sets.

In one dimension, it is well known that the Chebyshev coefficients c_k of a function $f(x)$ analytic on a neighbourhood of $[-1, 1]$ decay exponentially, $|c_k| \lesssim_\varepsilon \exp(-\alpha\, k)$, and the asymptotic decay rate α (cf. Definition 2.1.9) is equal to the parameter α of the largest Bernstein ellipse

$$E(\alpha) := \Big\{ \cosh(\tilde\alpha)\, \cos(\theta) + i\, \sinh(\tilde\alpha)\, \sin(\theta)\big) \mid \tilde\alpha \in [0, \alpha), \theta \in [0, 2\pi)\Big\} \qquad (4.12)$$

which can be inscribed into the domain of analyticity of f. In two dimensions, we have two decay rates α_1, α_2, and in the case of the conductivity function F_ζ we have two sets of singularities $\mathcal{S}_{\mathrm{temp}}$, $\mathcal{S}_{\mathrm{relax}}$ limiting the possible values of α_1 and α_2. This suggests that we partition the space of parameters ζ into *relaxation-constrained*, *mixed-constrained*, and *temperature-constrained* regimes depending on whether two, one, or zero of the decay rates are constrained by the singularities $\mathcal{S}_{\mathrm{relax}}$ rather than $\mathcal{S}_{\mathrm{temp}}$. In Section 4.4, we will characterize these parameter regimes more precisely and present asymptotic estimates regarding the number of significant Chebyshev coefficients in each case, a summary of which is provided in Table 4.1. We see that for fixed η, the cost of Algorithm 4.1 gradually increases from $\mathcal{O}\big(\eta^{-3/2}\big)$ to $\mathcal{O}\big(\beta^2\big)$ for increasing inverse temperature β which renders conductivity calculations at low temperatures (i.e., large β) particularly expensive. In Section 4.5, we present an alternative algorithm based on a pole expansion of F_ζ which provably reduces the cost of evaluating the local conductivity to $\mathcal{O}\big(\beta^{1/2}\, \eta^{-5/4}\big)$ inner products for all $\beta \gtrsim \eta^{-1/2}$ and whose actual scaling was empirically found to be $\mathcal{O}\big(\beta^{1/2}\, \eta^{-1.05}\big)$ inner products.

4.3 Aside: Why Tensor-Product Chebyshev Approximation?

Given the discussion in Chapter 2, the reader might be surprised to find that the focus in this chapter is mostly on polynomial rather than rational approximation. This section will provide some motivation for why we expect rational approximation of the conductivity to be ineffective, but before doing so we would like to point out an example where the expectations from Chapter 2 are satisfied.

Example 4.3.1 The singularities of the factor $f_{\text{temp}}(E_1, E_2)$ from (4.9) are decoupled in the sense that the locations $\mathcal{S}_{\beta,E_F}$ of the singularities in E_1 do not depend on the value of E_2 and vice versa. This special structure allows us to approximate $f_{\text{temp}}(E_1, E_2)$ by applying the one-dimensional theory from Chapter 2 to each of the variables separately, and since the one-dimensional singularities $\mathcal{S}_{\beta,E_F}$ are exactly those of the Fermi-Dirac function $f_{\beta,E_F}(E)$, we conclude from the comparison of Theorems 2.3.12 and 2.4.5 that rational approximation indeed performs much better at approximating $f_{\text{temp}}(E_1, E_2)$ than polynomial approximation.

The effectiveness of rational approximation to $f_{\beta,E_F}(E)$ and $f_{\text{temp}}(E_1, E_2)$ may heuristically be explained by the observation that rational functions allow us to concentrate computational effort in regions where the approximand lacks smoothness, cf. Figure 2.9a, and for both of the aforementioned functions this region is very small. Unfortunately, this property fails to hold for the factor $f_{\text{relax}}(E_1, E_2)$ from (4.10): assuming ω is sufficiently small, the region $E_2 - \omega + [-\eta, \eta]$ in E_1 where $f_{\text{relax}}(E_1, E_2)$ lacks smoothness moves through the entire domain $[-1, 1]$ of E_1 as E_2 varies over its domain $[-1, 1]$, and similarly with the roles of E_1 and E_2 interchanged. The ability of rational functions to direct computational effort to particular regions is thus worthless in the case of $f_{\text{relax}}(E_1, E_2)$ since the entire domain of approximation requires high resolution, and this property carries over to the conductivity function $F_\zeta(E_1, E_2)$.

The above heuristic motivates why we expect rational functions to offer no substantial benefit for approximating the conductivity function $F_\zeta(E_1, E_2)$ compared to polynomials, but it also raises the question of why we insist on approximating in the variables E_1, E_2 rather than, e.g., $x = E_1 - E_2$ and $y = E_1 + E_2$ such that $f_{\text{relax}}(E_1, E_2) = \frac{1}{x + \omega + \iota\eta}$ is a one-dimensional rational function in $\mathcal{R}_{01}$ in terms of x. To answer this question, let us reconsider the derivation of equation (4.8). In order to replace the eigenvalues of H with functions of H, it was crucial that we approximated the conductivity function $F_\zeta(E_1, E_2)$ in a tensor-product basis in E_1

and E_2. Had we approximated $F_\zeta(E_1, E_2)$ in terms of the variables x, y suggested above instead, the substitutions in (4.8) would not have worked out and hence it would have been unclear how to evaluate $F_\zeta(E_1, E_2)$ with a matrix argument H without computing the eigendecomposition of H.

We would like to conclude this aside by emphasising that the above discussion does not preclude the existence of more efficient local conductivity algorithms based on polynomial and rational approximations. Indeed, the pole expansion proposed in Section 4.5 is exactly the two-dimensional analogue of the rational approximation scheme presented in Theorem 2.4.1, and this scheme was found to be suboptimal in the one-dimensional case; hence it is likely that its two-dimensional extension is suboptimal as well. It remains an open problem whether there exists a rational approximation to the conductivity function which allows us to simultaneously exploit both the sparsity of the Chebyshev coefficients of the factor $f_{\mathrm{relax}}(E_1, E_2)$ discussed in Section 4.4 and the effectiveness of rational approximation for dealing with the Fermi-Dirac poles $\mathcal{S}_{\beta,E_F}$ as discussed in Section 2.4.

4.4 Chebyshev Coefficients of the Conductivity Function

Let us denote by α_{relax} the parameter of the ellipse penetrating the line $\omega + i\eta + [-1, 1]$ up to the endpoints, and by y_ζ half the width of this ellipse $E(\alpha_{\mathrm{relax}})$ along the line $\{z \mid \mathrm{Re}(z) = E_F\}$; see Figure 4.2. The partition into temperature-, mixed-, and relaxation-constrained conductivity parameters depends on whether and to what extent the Fermi-Dirac poles $\mathcal{S}_{\beta,E_F} = \big\{E_F + \frac{i\pi k}{\beta} \mid k \text{ odd}\big\}$ penetrate this ellipse $E(\alpha_{\mathrm{relax}})$.

- Relaxation-constrained: $\beta \in \big(0, \frac{\pi}{y_\zeta}\big]$. The Fermi-Dirac poles do not penetrate $E(\alpha_{\mathrm{relax}})$.
- Mixed-constrained: $\beta \in \big[\frac{\pi}{y_\zeta}, \frac{\pi}{\eta}\big]$. The Fermi-Dirac poles penetrate $E(\alpha_{\mathrm{relax}})$ but do not extend beyond the line $\omega + i\eta + [-1, 1]$.
- Temperature-constrained: $\beta \in \big[\frac{\pi}{\eta}, \infty\big)$. The Fermi-Dirac poles penetrate $E(\alpha_{\mathrm{relax}})$ beyond the line $\omega + i\eta + [-1, 1]$.

This partition (illustrated in Figure 4.2) allows us to formulate the following result.

Theorem 4.4.1 *There exist $\alpha_{\mathrm{diag}}(\zeta)$ and $\alpha_{\mathrm{anti}}(\zeta) > 0$ such that the Chebyshev coefficients $c_{k_1 k_2}$ of F_ζ are bounded by*

$$|c_{k_1,k_2}| \le C(\zeta) \exp\big[-\alpha_{\mathrm{diag}}(\zeta)\,(k_1 + k_2) - \alpha_{\mathrm{anti}}(\zeta)\,|k_1 - k_2|\big] \tag{4.13}$$

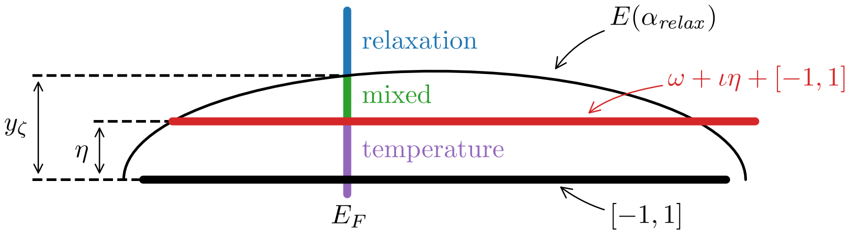


Figure 4.2: Partitioning of the conductivity parameters ζ depending on the location of the Fermi-Dirac poles (see Section 4.4).

for some $C(\zeta) < \infty$ independent of k_1, k_2. In the limit $\beta \to \infty$, $\omega, \eta \to 0$ with $|\omega| \lesssim \eta$ and assuming $E_F \in (-1, 1)$, we have that $y_\zeta \sim \sqrt{\eta}$,

$$\alpha_{\text{diag}}(\zeta) \sim \begin{cases} \eta & \text{if } \zeta \text{ is relaxation- or mixed-constrained,} \\ \beta^{-1} & \text{if } \zeta \text{ is temperature-constrained,} \quad \text{and} \end{cases}$$

$$\alpha_{\text{anti}}(\zeta) \sim \begin{cases} \sqrt{\eta} & \text{if } \zeta \text{ is relaxation-constrained,} \\ \beta^{-1} & \text{if } \zeta \text{ is mixed- or temperature-constrained.} \end{cases}$$

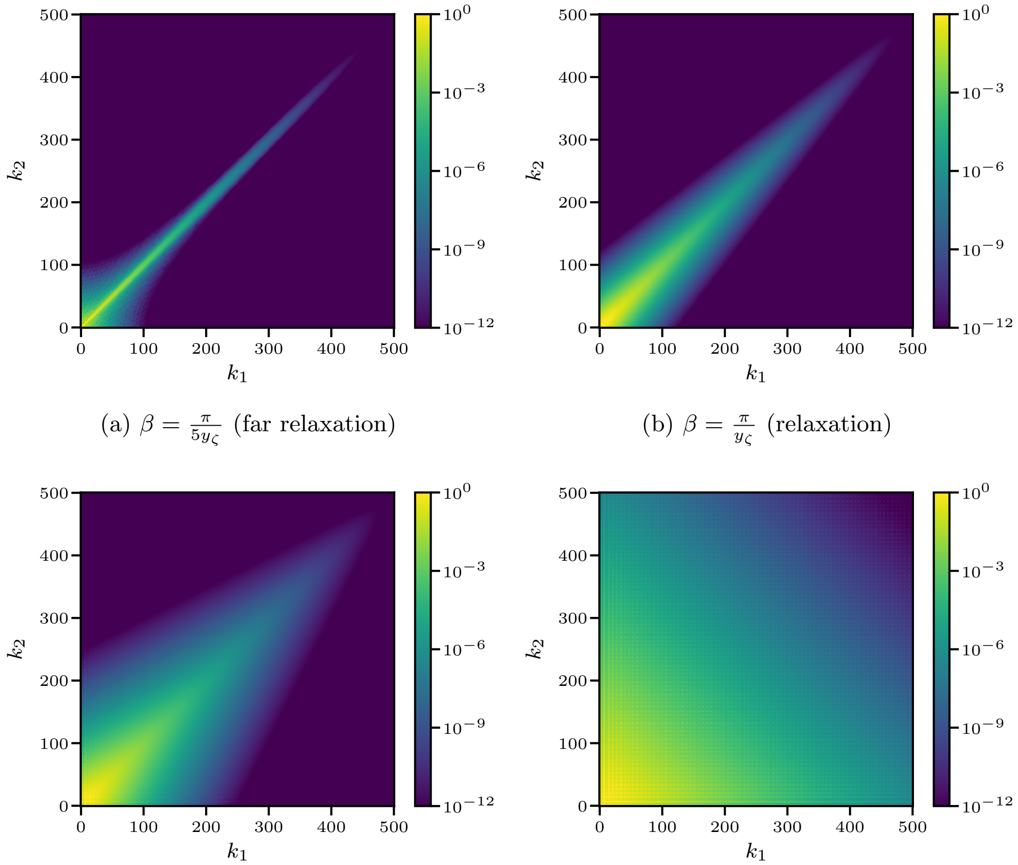
The notation $f(x) \sim g(x)$ is defined in Appendix A.2.

A proof of Theorem 4.4.1 and exact formulae for α_{relax}, α_{diag} and α_{anti} are provided in Section 4.7. Figures 4.3b to 4.3d show Chebyshev coefficients matching the predictions of Theorem 4.4.1 perfectly, and we note that Table 4.1 follows easily from Theorem 4.4.1.

We numerically observed the bound (4.13) to describe the correct decay behavior and the decay rates of $\alpha_{\text{diag}}(\zeta)$ and $\alpha_{\text{anti}}(\zeta)$ to be quantitatively accurate for temperature- and mixed-constrained parameters as well for relaxation-constrained parameters with β close to the critical value $\frac{\pi}{y_\zeta}$. For relaxation-constrained parameters far away from this critical value, however, the level lines of $c_{k_1 k_2}$ are piecewise concave rather than piecewise straight as predicted by Theorem 4.4.1, see Figure 4.3a, and this extra concentration reduces the number of significant Chebyshev coefficients from $\mathcal{O}\big(\eta^{-3/2}\big)$ to $\mathcal{O}\big(\eta^{-6/5}\big)$, see Figure 4.4. Since we do not have an explanation for this phenomenon, we will continue with the theoretically asserted scaling of $\mathcal{O}\big(\eta^{-3/2}\big)$ for clarity of exposition.

Theorem 4.4.1 suggests that we truncate the Chebyshev series in (4.6) using

$$K(\tau) := \Big\{(k_1, k_2) \in \mathbb{N}^2 \mid \exp\big(-\alpha_{\text{diag}}\,|k_1 + k_2| - \alpha_{\text{anti}}\,|k_1 - k_2|\big) \geq \tau\Big\},$$



(a) $\beta = \frac{\pi}{5y_\zeta}$ (far relaxation)

(b) $\beta = \frac{\pi}{y_\zeta}$ (relaxation)

(c) $\beta = \frac{\pi}{2\eta}$ (mixed)

(d) $\beta = \frac{2\pi}{\eta}$ (temperature)

Figure 4.3: Normalized Chebyshev coefficients $\hat{c}_{k_1 k_2} := |c_{k_1 k_2}|/|c_{00}|$ of the conductivity function F_ζ with $E_F = \omega = 0$, $\eta = 0.06$, and β as indicated.

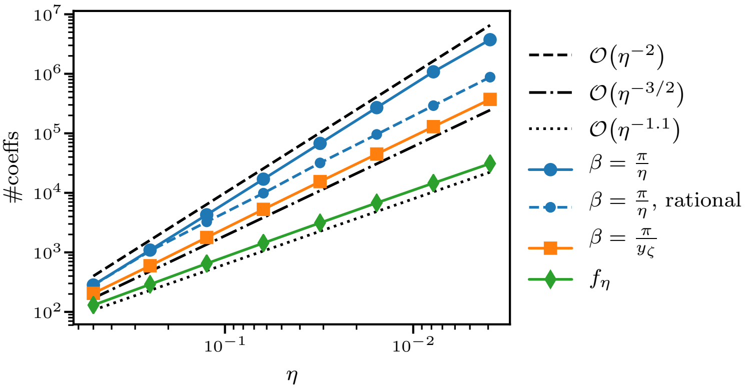


Figure 4.4: Number of normalized Chebyshev coefficients $\hat{c}_{k_1 k_2} := |c_{k_1 k_2}|/|c_{00}|$ larger than 10^{-3} for F_ζ with $E_F = \omega = 0$ and $f_\eta(E_1, E_2) := \frac{1}{E_1 - E_2 + i\eta}$. The "rational" line refers to the total number of Chebyshev coefficients in the pole expansion from Theorem 4.5.1 as described in Figure 4.5.

where here and in the following we no longer explicitly mention the dependence of $\alpha_{\text{diag}}(\zeta), \alpha_{\text{anti}}(\zeta)$ on ζ. The following theorem analyzes the error incurred by this approximation.

Theorem 4.4.2 *We have that*

$$\|F_\zeta - \tilde{F}_{\zeta,\tau}\|_{[-1,1]^2} \lesssim \alpha_{\text{diag}}^{-1}\, \alpha_{\text{anti}}^{-1}\, \tau\, |\log(\tau)|. \tag{4.14}$$

Proof. See Subsection 4.8.1. □

The $|\log(\tau)|$-factor in (4.14) varies very little over a large range of τ such that one may approximate it by a constant without losing much in accuracy. Doing so yields that we need to choose the truncation tolerance $\tau_\varepsilon := \alpha_{\text{diag}}\, \alpha_{\text{anti}}\, \varepsilon$ to guarantee an error $\|F_\zeta - F_{\zeta,\tau}\|_{[-1,1]^2} \lesssim \varepsilon$, and thus

$$|K(\tau_\varepsilon)| = \mathcal{O}\left(\frac{|\log(\alpha_{\text{diag}}\, \alpha_{\text{anti}}\, \varepsilon)|^2}{\alpha_{\text{diag}}\, \alpha_{\text{anti}}}\right).$$

4.5 Pole Expansion for Low-Temperature Calculations

We have seen in the previous subsection that for increasing β, the sparsity in the Chebyshev coefficients of F_ζ induced by the factor $\frac{1}{E_1 - E_2 + \omega + i\eta}$ decreases and the number of coefficients eventually scales as $\mathcal{O}(\beta^2)$; hence Algorithm 4.1 becomes expensive at low temperatures. To avoid this poor low-temperature scaling, we propose to expand F_ζ into a sum over the poles in $\mathcal{S}_{\text{temp}}$ as described in Theorem 4.5.1

below and apply Algorithm 4.1 to each term separately.

Theorem 4.5.1 *Let $k \in \mathbb{N}$ and denote by α_{k,β,E_F} the parameter of the ellipse through the poles $E_F \pm \frac{(2k+1)\,\pi i}{\beta}$ of the Fermi-Dirac function. There exists a function $R_{k,\beta,E_F}(E_1,E_2)$ analytic on the biellipse $E\big(\alpha_{k,\beta,E_F}\big)^2 \supset E\big(\alpha_{0,\beta,E_F}\big)^2$ such that*

$$F_\zeta(E_1,E_2) = \tfrac{1}{E_1-E_2+\omega+i\eta}\left(\sum_{z\in Z_k} \tfrac{1}{\beta}\,\tfrac{1}{(E_1-z)\,(E_2-z)} + R_{k,\beta,E_F}(E_1,E_2)\right), \tag{4.15}$$

where

$$Z_k := \big\{E_F + \tfrac{\ell\pi i}{\beta} \mid \ell \in \{-2k+1,-2k+3,\ldots,2k-3,2k-1\}\big\} \subset \mathcal{S}_{\beta,E_F}.$$

Proof. See Subsection 4.8.2. □

For k large enough, the remainder term (the last term in (4.15)) becomes relaxation constrained such that applying Algorithm 4.1 to this term becomes fairly efficient. For the pole terms, on the other hand, we propose to employ Algorithm 4.1 using the weighted Chebyshev approximation

$$\frac{1}{(E_1-z)\,(E_2-z)\,(E_1-E_2+\omega+i\eta)} \approx \sum_{k_1k_2\in K_z} c(z)_{k_1k_2}\,\frac{T_{k_1}(E_1)}{E_1-z}\,\frac{T_{k_2}(E_2)}{E_2-z},$$

where the weight $(E-z)^{-1}$ is chosen such that the two factors $(E_1-z)^{-1}$ and $(E_2-z)^{-1}$ on the left- and right-hand side cancel. The coefficients $c(z)_{k_1k_1}$ are therefore again the Chebyshev coefficients of a relaxation-constrained function

$$\frac{1}{E_1-E_2+\omega+i\eta} \approx \sum_{k_1k_2\in K_z} c(z)_{k_1k_2}\,T_{k_1}(E_1)\,T_{k_2}(E_2)$$

and exhibit the concentration described in Theorem 4.4.1. This leads us to the following algorithm where

$$w_{i_1i_2} := \langle\psi_{i_1}|M_a^{\mathrm{loc}}|\psi_{i_2}\rangle\,\langle\psi_{i_2}|M_b^{\mathrm{loc}}|e_{\ell,0,\alpha_\ell}\rangle\langle e_{\ell,0,\alpha_\ell}|\psi_{i_1}\rangle,$$

see (4.4).

Algorithm 4.2 Local conductivity via pole expansion

```
1: $\tilde\sigma_\ell(u) := \sum_{i_1,i_2} \frac{R_{k,\beta,E_F}(\varepsilon_{i_1},\varepsilon_{i_2})}{\varepsilon_{i_1}-\varepsilon_{i_2}+\omega+i\eta}\, w_{i_1,i_2}$, evaluated using Algorithm 4.1.
2: for $z \in Z_k$ do
3:     $\tilde\sigma_\ell(u) := \tilde\sigma_\ell(u) + \frac{1}{\beta}\sum_{i_1,i_2} \frac{w_{i_1 i_2}}{(\varepsilon_{i_1}-z)\,(\varepsilon_{i_2}-z)\,(\varepsilon_{i_1}-\varepsilon_{i_2}+\omega+i\eta)}$, evaluated using
       Algorithm 4.1 with the weighted Chebyshev polynomials $(E-z)^{-1}\,T_k(E)$.
4: end for
```

Theorem 4.5.2 *The dominant computational cost of Algorithm* 4.2 *is*

$$\#\mathrm{IP} = \mathcal{O}\big(k\,\eta^{-3/2}\big) + \begin{cases} \mathcal{O}\big(\eta^{-3/2}\big) & \text{if } \beta\,\eta^{1/2} \lesssim k, \\ \mathcal{O}\big(\frac{\beta\eta^{-1}}{k}\big) & \text{if } \beta\eta \lesssim k \lesssim \beta\,\eta^{1/2}, \\ \mathcal{O}\big(\frac{\beta^2}{k^2}\big) & \text{if } k \lesssim \beta\eta, \end{cases} \tag{4.16}$$

inner products if we assume that solving a single linear system of the form $(H_{\mathrm{loc}} - z)^{-1}\,v$ *is less expensive than* $\mathcal{O}\big(\eta^{-3/2}\big)$ *inner products (see Remark* 4.5.4*). This cost is minimized if we choose*

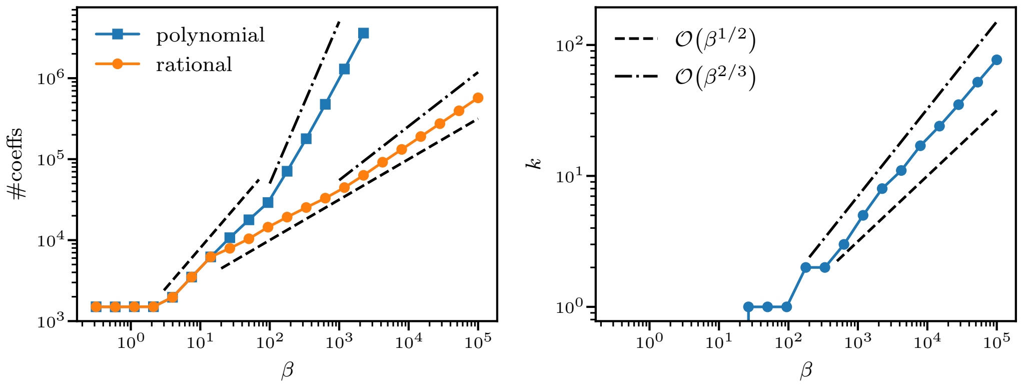
$$k \sim \begin{cases} 1 & \text{if } \beta \lesssim \eta^{-1/2}, \\ \beta^{1/2}\,\eta^{1/4} & \text{if } \eta^{-1/2} \lesssim \beta \lesssim \eta^{-3/2}, \\ \beta^{2/3}\,\eta^{1/2} & \text{if } \eta^{-3/2} \lesssim \beta, \end{cases} \tag{4.17}$$

which yields

$$\#\mathrm{IP} = \begin{cases} \mathcal{O}\big(\eta^{-3/2}\big) & \text{if } \beta \lesssim \eta^{-1/2}, \\ \mathcal{O}\big(\beta^{1/2}\,\eta^{-5/4}\big) & \text{if } \eta^{-1/2} \lesssim \beta \lesssim \eta^{-3/2}, \\ \mathcal{O}\big(\beta^{2/3}\,\eta^{-1}\big) & \text{if } \eta^{-3/2} \lesssim \beta. \end{cases}$$

Proof. It follows from Theorem 4.4.1 that the first term in (4.16) describes the cost of the for-loop in Algorithm 4.2 while the second term describes the cost of line 1. Since the first term is strictly increasing while the second is decreasing, the sum of the two terms is minimized by the unique k such that the first term equals the second term which one can readily verify to be given by (4.17) □

We note that Algorithm 4.2 reduces to Algorithm 4.1 if $\beta \lesssim \eta^{-1/2}$, but scales better than Algorithm 4.1 for larger values of β, e.g., for $\beta \sim \eta^{-1} \sim \chi$ we have $\#\mathrm{IP} = \mathcal{O}\big(\chi^{7/4}\big)$ in the case of Algorithm 4.2 while $\#\mathrm{IP} = \mathcal{O}\big(\chi^2\big)$ for Algorithm 4.1. The first term in (4.15) further reduces to $\mathcal{O}\big(k\,\eta^{-1.1}\big)$ if we assume the improved $\mathcal{O}\big(\eta^{-1.1}\big)$-scaling for the number of significant Chebyshev coefficients of $f(E_1,E_2) =$



(a) Number of coefficients

(b) Number of removed poles

Figure 4.5: (a) Number of normalized Chebyshev coefficients $\hat{c}_{k_1k_2} := |c_{k_1k_2}|/|c_{00}|$ larger than 10^{-3} for F_ζ with $\eta = 0.06$ and $E_F = \omega = 0$. The "polynomial" line counts the number of significant coefficients in the Chebyshev expansion from (4.6), while the "rational" line counts the sum of the Chebyshev coefficients of all the terms in the pole expansion from (4.15). The dashed lines denote $\mathcal{O}(\beta)$ and $\mathcal{O}(\beta^{1/2})$, respectively, and the dash-dotted lines denote $\mathcal{O}(\beta^2)$ and $\mathcal{O}(\beta^{2/3})$, respectively, cf. (4.18). (b) Index k for the set of poles Z_k from Theorem 4.5.1. This number was determined by increasing k starting from 0 until the number of coefficients reported in (a) stopped decreasing.

$\frac{1}{E_1 - E_2 + \omega + i\eta}$ suggested by Figure 4.4. In this case, the optimal choice of k and the corresponding costs are

$$k \sim \begin{cases} 1 \\ \beta^{1/2}\,\eta^{0.05} \\ \beta^{2/3}\,\eta^{0.37} \end{cases} \quad \text{and} \quad \#\text{IP} = \begin{cases} \mathcal{O}\big(\eta^{-1.1}\big) & \text{if } \beta \lesssim \eta^{-1/2}, \\ \mathcal{O}\big(\beta^{1/2}\,\eta^{-1.05}\big) & \text{if } \eta^{-1/2} \lesssim \beta \lesssim \eta^{-3/2}, \\ \mathcal{O}\big(\beta^{2/3}\,\eta^{-0.73}\big) & \text{if } \eta^{-3/2} \lesssim \beta. \end{cases} \tag{4.18}$$

These predictions are compared against numerical results in Figure 4.5 where we observe good qualitative agreement between the theory and the experiment. For $\beta \sim \eta^{-1} \sim \chi$, equation (4.18) yields $\#\text{IP} = \mathcal{O}\big(\chi^{1.55}\big)$ which is only marginally more expensive than the $\mathcal{O}\big(\chi^{1.5}\big)$ cost of Algorithm 4.1 in the case of relaxation-constrained parameters $\beta^2 \sim \eta^{-1} \sim \chi$. This is empirically demonstrated by the "rational" line in Figure 4.4.

Remark 4.5.3 Instead of running Algorithm 4.1 for each pole $z \in Z_k$ separately, we can apply Algorithm 4.1 to a group of poles $\tilde{Z} \subset Z_k$ if we weight the Chebyshev polynomials $T_k(E)$ with $q(E) := \prod_{z \in \tilde{Z}} (E - z)^{-1}$, and the same idea can also be used to improve the concentration of the Chebyshev coefficients of R_{k,β,E_F}. Grouping the poles in this manner reduces the computational cost of Algorithm 4.2, but amplifies the round-off errors by a factor $r := \max_{E \in [-1,1]} |q(E)| / \min_{E \in [-1,1]} |q(E)|$ such that the result is fully dominated by round-off errors if this ratio exceeds 10^{16}. Since $|q(E_F)| \sim \beta^{\#\tilde{Z}}$ while $|q(\pm 1)| \sim 1$, this means that we have to keep the group size rather small (e.g. $\#\tilde{Z} \leq 4$ for $\beta = 10^4$) to maintain numerical stability. We therefore conclude that grouping poles reduces the prefactor, but does not change the asymptotics of the computational cost of Algorithm 4.2.

Remark 4.5.4 We will see in Subsection 4.8.3 that the radius r of the local configurations $y^r_{\ell,\alpha_\ell}(A_{T(\ell)}\, u)$ must grow linearly with the maximal degree $k_{\max} := \max\big\{k_1 + k_2 \mid (k_1, k_2) \in K(\tau)\big\}$ of the polynomial approximation from (4.6) to achieve a constant error for all $k_{\max}$, and according to Theorem 4.4.1 the asymptotic scaling of $k_{\max}$ is given by

$$k_{\max} = \begin{cases} \mathcal{O}\big(\eta^{-1}\big) & \text{if } \zeta \text{ is relaxation- or mixed-constrained}, \\ \mathcal{O}\big(\beta\big) & \text{if } \zeta \text{ is temperature-constrained}. \end{cases}$$

From Table 1.1, it follows that solving a linear system $(H_{\text{loc}} - z)^{-1}\, v$ with $H_{\text{loc}} \in$

$\mathbb{R}^{m\times m}$ using a direct solver requires

$$\begin{aligned}\mathcal{O}\big(m^{3/2}\big) = \mathcal{O}\big(r^3\big) = \mathcal{O}\big(k_{\max}^3\big) \\ = \begin{cases}\mathcal{O}\big(\eta^{-3}\big) & \text{if } \zeta \text{ is relaxation- or mixed-constrained,} \\ \mathcal{O}\big(\beta^3\big) & \text{if } \zeta \text{ is temperature-constrained,}\end{cases}\end{aligned}$$

floating-point operations, while approximating $p(E) \approx 1/(E-z)$ and evaluating $p(H_{\text{loc}}) \approx (H_{\text{loc}} - z)^{-1}$ (or equivalently, using an iterative linear solver like conjugate gradients) requires

$$\mathcal{O}\big(\text{degree}(p)\, m\big) = \begin{cases}\mathcal{O}\big(\beta\, \eta^{-2}\big) & \text{if } \zeta \text{ is relaxation- or mixed-constrained,} \\ \mathcal{O}\big(\beta^3\big) & \text{if } \zeta \text{ is temperature-constrained}\end{cases}$$

floating-point operations where we used that $\text{degree}(p) = \mathcal{O}\big(|\operatorname{Im}(z)|^{-1}\big) = \mathcal{O}\big(\beta\big)$ according to Theorem 2.3.8. We conclude that iterative solvers scale slightly better than direct ones in the relaxation- and mixed-constrained cases and scale as well as direct ones in the temperature-constrained case.

Similarly, we find that the cost of computing $\mathcal{O}\big(\eta^{-3/2}\big)$ inner products is

$$\mathcal{O}\big(\eta^{-3/2}\, m\big) = \begin{cases}\mathcal{O}\big(\eta^{-7/2}\big) & \text{if } \zeta \text{ is relaxation- or mixed-constrained,} \\ \mathcal{O}\big(\eta^{-3/2}\, \beta^2\big) & \text{if } \zeta \text{ is temperature-constrained,}\end{cases}$$

floating-point operations; hence the assumption in Theorem 4.5.2 is satisfied if $\beta \lesssim \eta^{-3/2}$.

4.6 Remarks Regarding Implementation

This section remarks on a few practical aspects of the above algorithms.

4.6.1 Memory Requirements

Algorithm 4.1 as formulated above suggests that we precompute and store both the vectors $|v_{k_1}\rangle$ for all $k_1 \in K_1$ and $|w_{k_2}\rangle$ for all $k_2 \in K_2$, but this requires more memory than needed since we can rewrite the algorithm as follows.

Algorithm 4.3 Memory-optimised version of Algorithm 4.1

```
Precompute $|v_{k_1}\rangle$ for all $k_1 \in K_1$ as in Algorithm 4.1.
for $k_2 \in K_2$ in ascending order do
    Evaluate $|w_{k_2}\rangle$ using the recurrence relation (4.7).
    Discard $|w_{k_2-2}\rangle$ as it will no longer be needed.
    Compute the inner products $\langle v_{k_1}|w_{k_2}\rangle$ for all $k_1$ such that $(k_1, k_2) \in K$, and
      accumulate the results as in Algorithm 4.1.
end for
```

Furthermore, even caching all the vectors $|v_{k_1}\rangle$ is not needed if the function to be evaluated is relaxation-constrained: it follows from the wedge-like shape of the Chebyshev coefficients of such functions shown in Figure 4.3b that in every iteration of the loop in Algorithm 4.3, we only need vectors $|v_{k_1}\rangle$ with index k_1 within some fixed distance from k_2. The vectors $|v_{k_1}\rangle$ can hence be computed and discarded on the fly just like $|w_{k_2}\rangle$, albeit with a larger lag between computing and discarding. Quantitatively, this reduces the memory requirements from $\mathcal{O}\big(\eta^{-1}\, m\big)$ for both Algorithms 4.1 and 4.3 to $\mathcal{O}\big(\eta^{-1/2}\, m\big)$ for the final version described above, assuming the function to be evaluated is relaxation-constrained.

4.6.2 Choosing the Approximation Scheme

Algorithms 4.1 and 4.2 involve three basic operations, namely matrix-vector products, inner products and linear system solves, and a fundamental assumption in their derivation was that matrix-vector and inner products are approximately equally expensive and linear system solves are not significantly more expensive than that (see Theorem 4.5.2 for the precise condition). The former assumption is true in the sense that both matrix-vector and inner products scale linearly in the matrix size m, but their prefactors are very different: the inner product $\langle w \,|\, v\rangle$ requires $2m - 1$ floating-point operations, while the cost of the matrix-vector product $H\,|v\rangle$ is approximately equal to twice the number of nonzeros in H. Even in the simplest case of a single triangular lattice and a tight-binding Hamiltonian H involving only nearest-neighbour terms and s and p orbitals, the number of nonzeros per column of H is about 6 (number of neighbours) times 4 (number of orbitals); hence the cost of evaluating $H\,|v\rangle$ is approximately $48m$ which is 24 times more expensive than the inner product. Similarly, the assumption regarding the costs of linear system solves holds true in the asymptotic sense as discussed in Remark 4.5.4, but the situation may look very different once we include the prefactors. This observation has two practical implications.

- Rather than choosing the number of removed poles k in Theorem 4.5.1 solely to minimise the number of coefficients, one should benchmark the runtimes of inner products, matrix-vector products and linear system solves and choose the k which yields the smallest overall runtime.

- Fairly small values of η are required before the wedge shown in Figure 4.3b becomes thin enough that the savings due to a smaller number of inner products make a significant difference compared to the cost of the matrix vector products, and even smaller η are required to compensate for the additional costs of solving the linear systems introduced by the pole expansion in Theorem 4.5.1.

We have seen in Remark 4.5.4 that the matrix size m must scale with η^{-2} in order to achieve a constant error in the local conductivities σ_{loc}; hence the latter point implies that demonstrating the savings brought about by the sparsity of the Chebyshev coefficients in a physically meaningful setting requires large-scale computations which are beyond the scope of this thesis and will be the topic of future work.

4.7 Proof of Theorem 4.4.1

4.7.1 Approximation theory background

This subsection briefly recalls some concepts from approximation theory and introduces the notation used in the remainder of this section. A textbook introduction to the topics discussed here can be found, e.g., in [Tre13].

Joukowsky map $\phi(z)$. The three-term recurrence relation (4.7) for the Chebyshev polynomials $T_k(x)$ implies the identity

$$T_k\big(\phi(z)\big) := \frac{z^k + z^{-k}}{2} \qquad \text{where} \qquad \phi(z) := \frac{z + z^{-1}}{2}$$

as one can easily verify by induction, and the Bernstein ellipses (4.12) can be expressed in terms of the Joukowsky map $\phi(z)$ as

$$E(\alpha) = \{\phi(z) \mid z \in \mathbb{C},\, 0 \le \log|z| < \alpha\}.$$

Parameter function $\alpha_b(x)$. It will be convenient in the following to express $E(\alpha)$ in terms of the variable $x := \phi(z)$, which requires us to study the inverse Joukowsky map $\phi_\pm^{-1}(x) = x \pm \sqrt{x^2-1}$. Since $\phi(z) = \phi\big(z^{-1}\big)$, this inverse has two branches related by $\phi_\pm^{-1}(x) = \big(\phi_\mp^{-1}(x)\big)^{-1}$, and given any curve $b \subset \mathbb{C}$ connecting the two branch points $x = \pm 1$, we define

$$\phi_b^{-1}(x) := x + \sqrt[b]{x^2 - 1}$$

where $\sqrt[b]{x^2-1}$ denotes the branch of $\sqrt{x^2-1}$ with branch cut along b and sign such that $\phi_b^{-1}(\infty) = \infty$. The Bernstein ellipses $E(\alpha)$ then become the level sets

$$E(\alpha) = \{x \in \mathbb{C} \mid \alpha_{[-1,1]}(x) < \alpha\}$$

of the parameter function

$$\alpha_b(x) := \log |\phi_b^{-1}(x)|.$$

The following properties of $\alpha_b(x)$ follow immediately from the above discussion.

Lemma 4.7.1 *We have that*

- $\alpha_b(x) = 0$ *for all* $x \in [-1, 1]$ *and all branch cuts* b*,*
- $\alpha_{[-1,1]}(x) \geq 0$ *for all* $x \in \mathbb{C}$*,*
- $\alpha_b(x+0n) = -\alpha_b(x-0n)$ *for all* $x \in b$ *and all branch cuts* b*, where the notation* $x \pm 0n$ *indicates that we evaluate* $\alpha_b(x)$ *on different sides of the branch cut.*

We remark that $\alpha_{[-1,1]}(x)$ is in fact the Green's function $g_{[-1,1]}(x) = \operatorname{Re}\bigl(G_{[-1,1]}(x)\bigr)$ of the log-map $G_{[-1,1]}(x)$ from Theorem 2.3.4, and much of the discussion above is closely related to the material presented in Chapter 2.

Zero-width contours. As in Definition 2.0.2, we define $\partial\gamma$ for curves $\gamma \subset \mathbb{C}$ as the counterclockwise contour around a domain of infinitesimal width, e.g.,

$$\partial[-1, 1] = \bigl([-1, 1] + 0i\bigr) \cup \bigl([-1, 1] - 0i\bigr),$$

where the signed zero in the imaginary part indicates which branch to evaluate for a function with branch cut along $[-1, 1]$.

Example 4.7.2 Using the definition of $\partial[-1, 1]$ and $\sqrt[{[-1,1]}]{\cdot}$, we compute

$$\begin{aligned}\int_{\partial[-1,1]} \sqrt[{[-1,1]}]{x^2-1}\,dx &= \int_{1+0i}^{-1+0i} i\,\sqrt{1-x^2}\,dx + \int_{-1-0i}^{1-0i} (-i)\,\sqrt{1-x^2}\,dx \\ &= -2\,i \int_{-1}^{1} \sqrt{1-x^2}\,dx = -\pi\,i.\end{aligned}$$

The signs of $\sqrt[{[-1,1]}]{x^2-1}$ given on the first line follow from the requirement that $\sqrt[{[-1,1]}]{x^2-1} \to \infty$ for $x \to \infty$.

Exponential decay with asymptotic rate α. We recall the notations

$$a_k \leq_\varepsilon C(\alpha)\,\exp\bigl(-\alpha\,k\bigr) \qquad \text{and} \qquad a_k \lesssim_\varepsilon \exp\bigl(-\alpha\,k\bigr)$$

from Definition 2.1.9.

Analyticity in two dimensions. We extend the notion of analyticity to two-dimensional functions $f(z_1, z_2)$ as follows.

Definition 4.7.3 *A function $f : \Omega \to \mathbb{C}$ with $\Omega \subset \mathbb{C}^2$ is called* analytic *if $f(z_1, z_2)$ is analytic in the one-dimensional sense in each variable z_1, z_2 separately for every $(z_1, z_2) \in \Omega$.*

By a famous result due to Hartogs (see e.g. [Kra01, Theorem 1.2.5]), a function $f(z_1, z_2)$ analytic in the above one-dimensional sense is continuous and differentiable in the two-dimensional sense. Furthermore, it is known that if $f(z_1, z_2)$ is analytic on the biannulus $A(r_1) \times A(r_2)$ with $A(r) := \{z \mid r^{-1} < |z| < r\}$, it can be expanded into a Laurent series

$$f(z_1, z_2) = \sum_{k_1,k_2=-\infty}^{\infty} a_{k_1k_2}\, z_1^{k_1}\, z_2^{k_2}$$

with coefficients given by

$$a_{k_1k_2} = -\frac{1}{4\pi^2} \int_{\gamma_2} \int_{\gamma_1} f(z_1, z_2)\; z_1^{-k_1-1}\, z_2^{-k_2-1}\, dz_1\, dz_2$$

for any bicontour $\gamma_1 \times \gamma_2$ where $\gamma_\ell \subset A(r_\ell)$ are two closed contours winding once around the origin, see [Sch05, Theorem 1.5.26].

4.7.2 Auxiliary results

We next establish a contour-integral formula for the Chebyshev coefficients of analytic functions in Theorem 4.7.4 and demonstrate in Theorem 4.7.5 how this formula translates into a bound on the Chebyshev coefficients. Both results are straightforward generalizations of the one-dimensional results (see e.g., [Tre13]) except that we allow for a general branch cut in Theorem 4.7.5 which will be important in Subsection 4.7.3.

Theorem 4.7.4 *A function $f(x_1, x_2)$ analytic on $[-1, 1]^2$ can be expanded into a Chebyshev series*

$$f(x_1, x_2) = \sum_{k_1,k_2=0}^{\infty} c_{k_1k_2}\, T_{k_1}(x_1)\, T_{k_2}(x_2) \quad \text{on } [-1, 1]^2 \tag{4.19}$$

with coefficients $c_{k_1k_2}$ given by

$$c_{k_1k_2} = -\tfrac{(2-\delta_{k_10})(2-\delta_{k_20})}{4\pi^2} \int_{\partial[-1,1]} \int_{\partial[-1,1]} f(x_1,x_2)\, \frac{T_{k_1}(x_1)}{\sqrt[{[-1,1]}]{x_1^2-1}}\, \frac{T_{k_2}(x_2)}{\sqrt[{[-1,1]}]{x_2^2-1}}\, dx_1\, dx_2.$$

Proof. $f(x_1,x_2)$ is analytic on $[-1,1]^2$ and $\phi(z)$ maps the unit circle $\{|z| = 1\}$ holomorphically onto $[-1,1]$; thus $f\big(\phi(z_1),\phi(z_2)\big)$ is analytic on $\{|z| = 1\}^2$ and can be expanded into a Laurent series

$$f\big(\phi(z_1),\phi(z_2)\big) = \sum_{k_1,k_2=-\infty}^{\infty} a_{k_1,k_2}\, z_1^{k_1}\, z_2^{k_2} \tag{4.20}$$

with coefficients $a_{k_1k_2}$ given by

$$a_{k_1k_2} = -\frac{1}{4\pi^2} \int_{|z_2|=1} \int_{|z_1|=1} f\big(\phi(z_1),\phi(z_2)\big)\, z_1^{-k_1-1}\, z_2^{-k_2-1}\, dz_1\, dz_2. \tag{4.21}$$

Since $\phi(z) = \phi\big(z^{-1}\big)$, we conclude that $a_{k_1k_2}$ is symmetric about the origin in both k_1 and k_2, i.e., $a_{k_1,k_2} = a_{-k_1,k_2}$ and $a_{k_1,k_2} = a_{k_1,-k_2}$. The terms in (4.20) can therefore be rearranged as a Chebyshev series in $\phi(z_1)$ and $\phi(z_2)$,

$$\begin{aligned} f\big(\phi(z_1),\phi(z_2)\big) &= \sum_{k_1,k_2=0}^{\infty} (2-\delta_{k_10})(2-\delta_{k_20})\, a_{k_1k_2}\, \frac{z_1^{k_1}+z_1^{-k_1}}{2}\, \frac{z_2^{k_2}+z_2^{-k_2}}{2} \\ &= \sum_{k=0}^{\infty} c_{k_1k_2}\, T_{k_1}\big(\phi(z_1)\big)\, T_{k_2}\big(\phi(z_2)\big), \end{aligned}$$

which is (4.19) with $c_{k_1k_2} := (2-\delta_{k_10})(2-\delta_{k_20})\, a_{k_1k_2}$. The formula for the coefficients follows by substituting

$$z_\ell \to \phi_{[-1,1]}^{-1}(x_\ell), \qquad dz_\ell \to \frac{\phi(x_\ell)}{\sqrt[{[-1,1]}]{x^2-1}}\, dx_\ell \quad \text{and} \quad \{|z_\ell| = 1\} \to \partial[-1,1]$$

for both $\ell = 1$ and $\ell = 2$ in the integrals in (4.21). □

Theorem 4.7.5 *Let $\Omega_1, \Omega_2 \subseteq \mathbb{C}$ be two simply connected sets such that both sets contain -1 and 1. We then have that*

$$\begin{aligned} &\left| \frac{(2-\delta_{k_10})(2-\delta_{k_20})}{4\pi^2} \int_{\partial\Omega_2} \int_{\partial\Omega_1} f(x_1,x_2)\, \frac{T_{k_1}(x_1)}{\sqrt[b_1]{x_1^2-1}}\, \frac{T_{k_2}(x_2)}{\sqrt[b_2]{x_2^2-1}}\, dx_1\, dx_2 \right| \le \dots \\ &\qquad \le C(\partial\Omega_1)\, C(\partial\Omega_2)\, \|f\|_{\partial\Omega_1\times\partial\Omega_2}\, \exp\big(-\alpha_1 k_1 - \alpha_2 k_2\big) \end{aligned}$$

for all $k_1, k_2 \in \mathbb{N}$ *and all branch cuts* $\big(b_\ell \subset \Omega_\ell\big)_{\ell \in \{1,2\}}$ *connecting* $-1, 1$*, where*

$$\Big(\alpha_\ell := \min \alpha_{b_\ell}(\partial\Omega_\ell)\Big)_{\ell\in\{1,2\}} \qquad \textit{and} \qquad C(\partial\Omega) := \frac{1}{\pi} \int_{\phi_b^{-1}(\partial\Omega)} \frac{|dz|}{|z|}.$$

Proof. Reversing the substitutions in the proof of Theorem 4.7.4 transforms the expression on the left-hand side to (4.21) up to a factor of $(2 - \delta_{k_1 0})(2 - \delta_{k_2 0})$ and the integrals running over $\phi_b^{-1}(\partial\Omega_\ell)$ instead of $\{|z_\ell| = 1\}$ for $\ell \in \{1, 2\}$. The claim follows by bounding these integrals using Hölder's inequality. □

We illustrate the application of Theorems 4.7.4 and 4.7.5 by proving the following corollary which can be found e.g., in [BM48, Theorem 11], [Tre17, Lemma 5.1] and [Boy09, Theorem 11].

Corollary 4.7.6 *The Chebyshev coefficients of a function* $f(x_1, x_2)$ *analytic on* $E(\alpha_1) \times E(\alpha_2)$ *are bounded by*

$$|c_{k_1k_2}| \le_\varepsilon 4\,\|f\|_{\partial E(\alpha_1)\times\partial E(\alpha_2)} \exp\!\big(-\alpha_1 k_1 - \alpha_2 k_2\big) \qquad \textit{for all } k_1, k_2 \in \mathbb{N}. \tag{4.22}$$

Proof. $f(x_1, x_2)$ is analytic on $[-1, 1]^2 \subset E(\alpha_1) \times E(\alpha_2)$; thus Theorem 4.7.4 says that we can expand $f(x_1, x_2)$ into a Chebyshev series with coefficients given by

$$c_{k_1k_2} = -\tfrac{(2-\delta_{k_10})(2-\delta_{k_20})}{4\pi^2} \int_{\partial\Omega_2}\int_{\partial\Omega_1} f(x_1, x_2)\, \frac{T_{k_1}(x_1)}{{}^{[-1,1]}\!\sqrt{x_1^2 - 1}}\, \frac{T_{k_2}(x_2)}{{}^{[-1,1]}\!\sqrt{x_2^2 - 1}}\, dx_1\, dx_2$$

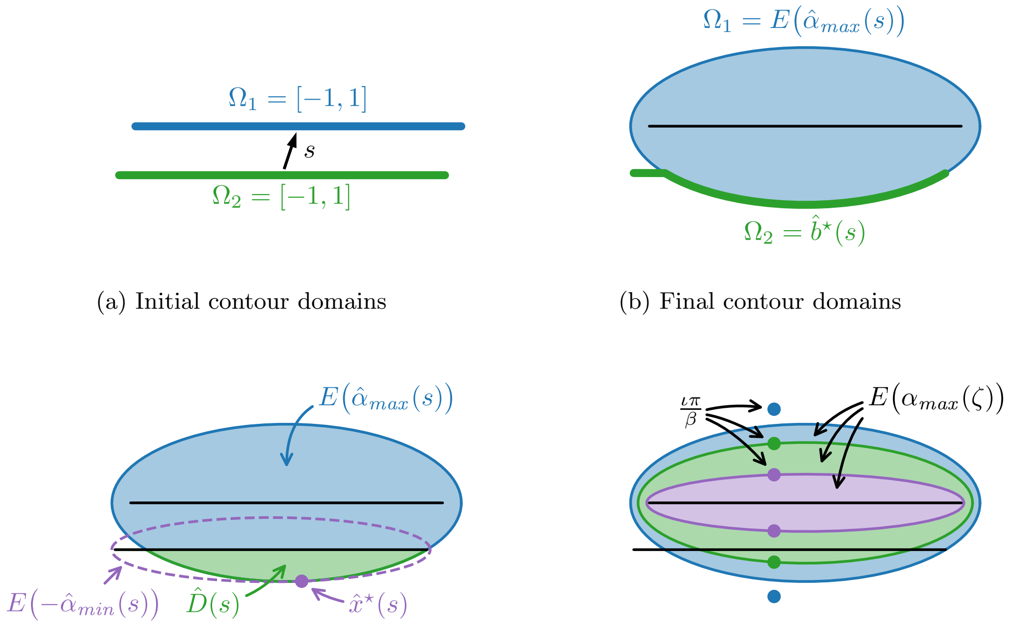
where $\Omega_1 = \Omega_2 = [-1, 1]$. Using Cauchy's integral theorem and the analyticity of $f(x_1, x_2)$, we can replace the two contour domains $\Omega_1 = \Omega_2 = [-1, 1]$ with $\Omega_\ell = E(\tilde\alpha_\ell)$ for any $\tilde\alpha_\ell < \alpha_\ell$, which by Theorem 4.7.5 implies

$$|c_{k_1,k_2}| \le 4\,\|f\|_{\partial E(\tilde\alpha_1)\times\partial E(\tilde\alpha_2)} \exp\!\big(-\tilde\alpha_1 k_1 - \tilde\alpha_2 k_2\big)$$

where we used $C\big(\partial E(\alpha)\big) = \frac{1}{\pi}\int_{|z|=\exp(\alpha)} \frac{|dz|}{|z|} = 2$ and $\alpha_{[-1,1]}\big(\partial E(\alpha)\big) = \alpha$. This is precisely the bound (4.22). □

4.7.3 Chebyshev coefficients of the conductivity function

This subsection establishes the bound (4.13) with explicit formulae for $\alpha_{\mathrm{diag}}(\zeta)$ and $\alpha_{\mathrm{anti}}(\zeta)$, which will be done in two steps. First, we will establish in Theorem 4.7.7 below a bound on the Chebyshev coefficients of the factor $f(x_1, x_2) = \frac{1}{x_1 - x_2 + s}$ from (4.10), where for notational convenience we set $s := \omega + i\eta$. The extension to the conductivity function F_ζ will then be an easy modification of Theorem 4.7.7.


$\Omega_1 = [-1,1]$
s
$\Omega_2 = [-1,1]$
$\Omega_1 = E\big(\hat\alpha_{max}(s)\big)$
$\Omega_2 = \hat b^\star(s)$
$E\big(\hat\alpha_{max}(s)\big)$
$E\big(-\hat\alpha_{min}(s)\big)$
$\hat D(s)$
$\hat x^\star(s)$
$\frac{\iota\pi}{\beta}$
$E\big(\alpha_{max}(\zeta)\big)$

(a) Initial contour domains

(b) Final contour domains

(c) Definitions

(d) $E\big(\alpha_{\max}(\zeta)\big)$ for three β

Figure 4.6: Illustration of the various definitions in Subsection 4.7.3.

We note that $x_1 \mapsto \frac{1}{x_1 - x_2 + s}$ is analytic at all $x_1 \in \mathbb{C}$ except $x_1 = x_2 - s$, and likewise $x_2 \mapsto \frac{1}{x_1 - x_2 + s}$ is analytic at all $x_2 \in \mathbb{C}$ except $x_2 = x_1 + s$. The condition that $\frac{1}{x_1 - x_2 + s}$ is analytic on a domain $\Omega_1 \times \Omega_2$ is thus equivalent to $\big(\Omega_1 + s\big) \cap \Omega_2 = \{\}$ which is clearly the case for $\Omega_1 = \Omega_2 = [-1,1]$ and $\operatorname{Im}(s) \neq 0$, see Figure 4.6a. By Theorem 4.7.4, we can thus expand $\frac{1}{x_1 - x_2 + s}$ into a Chebyshev series with coefficients given by

$$c_{k_1 k_2} = -\frac{(2 - \delta_{k_1 0})(2 - \delta_{k_2 0})}{4\pi^2} \int_{\partial\Omega_2} \int_{\partial\Omega_1} \frac{1}{x_1 - x_2 + s} \, \frac{T_{k_1}(x_1)}{\sqrt[b_1]{x_1^2 - 1}} \, \frac{T_{k_2}(x_2)}{\sqrt[b_2]{x_2^2 - 1}} \, dx_1 \, dx_2$$

where for now $\Omega_1 = \Omega_2 = b_1 = b_2 = [-1,1]$.

As in the proof of Corollary 4.7.6, we will next use Cauchy's integral theorem repeatedly to move the contour domains $\Omega_{1,2}$ to appropriate shapes and then employ Theorem 4.7.5 to bound the Chebyshev coefficients. To this end, let us introduce

$$\hat\alpha_{\max}(s) := \min\{\alpha_{[-1,1]}(\pm 1 - s)\} = \alpha_{[-1,1]}\big(1 - |\operatorname{Re}(s)| - i \operatorname{Im}(s)\big),$$

which is the parameter of the ellipse $E\big(\hat\alpha_{\max}(s)\big)$ penetrating the line $[-1,1] - s$ up to the endpoints ±1 + s, and let us denote by

$$\hat{D}(s) := \Big(E\big(\hat{\alpha}_{\max}(s)\big) + s\Big) \cap \big\{x \mid \operatorname{Im}(x) < 0\big\}$$

the portion of $E\big(\hat{\alpha}_{\max}(s)\big)+s$ penetrating $[-1,1]$ (see Figure 4.6c). Arguing similarly as above, we see that $\frac{1}{x_1-x_2+s}$ is analytic on $[-1,1] \times \big([-1,1] \cup \overline{\hat{D}(s)}\big)$; thus we can replace $\Omega_2 = [-1,1]$ with $\Omega_2 = [-1,1] \cup \hat{D}(s)$ without changing the value of the integral. We next move the branch cut $b_2 = [-1,1]$ to the lower boundary of Ω_2,

$$b_2 = \hat{b}^\star(s) := \big([-1,1] \setminus \hat{D}(s)\big) \cup \{x \in \partial\hat{D}(s) \mid \operatorname{Im}(s) < 0\},$$

which allows us to replace $\Omega_2 = [-1,1] \cup \hat{D}(s)$ with $\Omega_2 = \hat{b}^\star(s)$ and finally replace $\Omega_1 = [-1,1]$ with $\Omega_1 = E(\tilde{\alpha}_1)$ for any $\tilde{\alpha}_1 < \hat{\alpha}_{\max}(s)$, see Figure 4.6b. By Theorem 4.7.5, these final contours imply the bound

$$|c_{k_1k_2}| \lesssim_\varepsilon \exp\big(-\hat{\alpha}_{\max}(s)\, k_1 - \hat{\alpha}_{\min}(s)\, k_2\big) \tag{4.23}$$

with

$$\hat{\alpha}_{\min}(s) := \min \alpha_{\hat{b}^\star(s)}\big(\partial\hat{b}^\star(s)\big) = -\max \alpha_{[-1,1]}\big(\hat{b}^\star(s)\big), \tag{4.24}$$

where for the second equality we used Lemma 4.7.1. We note that the last expression in (4.24) may be interpreted as minus the parameter of the smallest ellipse containing $\hat{D}(s)$, see Figure 4.6c.

By the symmetry of $\frac{1}{x_1-x_2+s}$, the bound (4.23) also holds with the roles of k_1, k_2 interchanged, and since $\hat{\alpha}_{\max}(s) > 0$ but $\hat{\alpha}_{\min}(s) < 0$, we may summarize the two bounds with

$$|c_{k_1k_2}| \lesssim_\varepsilon \begin{cases} \exp\big(-\hat{\alpha}_{\max}(s)k_1 - \hat{\alpha}_{\min}(s)\, k_2\big) & \text{if } k_1 \geq k_2, \\ \exp\big(-\hat{\alpha}_{\min}(s)\, k_1 - \hat{\alpha}_{\max}(s)k_2\big) & \text{if } k_1 \leq k_2. \end{cases} \tag{4.25}$$

Rewriting (4.25) in the form (4.26), we arrive at the following theorem.

Theorem 4.7.7 *The Chebyshev coefficients* $c_{k_1k_2}$ *of* $f(x_1,x_2) := \frac{1}{x_1-x_2+s}$ *with* $\operatorname{Re}(s) \in [-1,1]$ *are bounded by*

$$|c_{k_1,k_2}| \lesssim_\varepsilon \exp\big(-\hat{\alpha}_{\mathrm{diag}}(s)\,(k_1+k_2) - \hat{\alpha}_{\mathrm{anti}}(s)\,|k_1-k_2|\big) \tag{4.26}$$

where

$$\hat{\alpha}_{\mathrm{diag}}(s) := \frac{1}{2}\Big(\hat{\alpha}_{\max}(s) + \hat{\alpha}_{\min}(s)\Big) \quad \textit{and} \quad \hat{\alpha}_{\mathrm{anti}}(s) := \frac{1}{2}\Big(\hat{\alpha}_{\max}(s) - \hat{\alpha}_{\min}(s)\Big).$$

A closer inspection of the above argument reveals that the bound (4.26) applies to any function $f(x_1,x_2) = \frac{g(x_1,x_2)}{x_1-x_2+s}$ as long as $g(x_1,x_2)$ is analytic on $E\big(\hat\alpha_{\max}(s)\big)^2$, and in particular it applies to the conductivity function $F_\zeta(E_1,E_2) = \frac{f_{\text{temp}}(E_1,E_2)}{E_1-E_2+\omega+i\eta}$ if the singularities $\mathcal{S}_{\text{temp}}$ of f_{temp} from (4.11) satisfy

$$E\big(\alpha_{\text{relax}}\big)^2 \cap \mathcal{S}_{\text{temp}} = \{\} \iff E\big(\alpha_{\text{relax}}\big) \cap \mathcal{S}_{\beta,E_F} = \{\} \iff \beta \le \frac{\pi}{y_\zeta},$$

i.e., if ζ is relaxation-constrained (note that $\alpha_{\text{relax}} = \hat\alpha_{\max}(\omega+i\eta)$). Furthermore, the argument and hence the bound (4.26) can be extended to the non-relaxation-constrained case $\beta \ge \frac{\pi}{y_\zeta}$ if we replace $\hat\alpha_{\max}(\zeta)$ with

$$\alpha_{\max}(\zeta) = \min\big\{\hat\alpha_{\max}(\omega+i\eta),\, \alpha_{[-1,1]}\big(E_F + \tfrac{\pi i}{\beta}\big)\big\},$$

see Figure 4.6d, and $\hat\alpha_{\min}(s)$ with $\alpha_{\min}(\zeta)$ defined analogously to $\hat\alpha_{\min}$ but starting from $\alpha_{\max}(\zeta)$ instead of $\hat\alpha_{\max}(s)$.

4.7.4 Asymptotics

To complete the proof of Theorem 4.4.1, it remains to show the asymptotic scaling of y_ζ, $\alpha_{\text{diag}}(\zeta)$ and $\alpha_{\text{anti}}(\zeta)$, which we will do using the following auxiliary result.

Lemma 4.7.8 *We have that*

$$\alpha_{[-1,1]}(x) \sim |\operatorname{Im}(x)| \qquad \text{for } x \to x^\star \text{ with } x^\star \in (-1,1), \tag{4.27}$$

$$\alpha_{[-1,1]}(x) \sim \sqrt{|x \mp 1|} \qquad \text{for } x \to \pm 1 \text{ with } \pm\operatorname{Re}(x) - 1 \ge C|\operatorname{Im}(x)|. \tag{4.28}$$

Proof. As noted after Lemma 4.7.1, we have that $\alpha_{[-1,1]}(x) = g_{[-1,1]}(x)$; hence (4.27) follows immediately from (2.17), and the derivation of (4.28) is analogous to that of (2.19) and (2.20) in Theorem 2.3.12. □

Lemma 4.7.8 immediately yields $\alpha_{\text{relax}} = \alpha_{[-1,1]}\big(1-|\omega|+i\eta\big) \sim \sqrt{\eta}$ which in turn implies $y_\zeta \sim \sqrt{\eta}$ as one can verify by Taylor-expanding the formula

$$y_\zeta = \sinh(\alpha_{\text{relax}})\sqrt{1 - \frac{E_F^2}{\cosh(\alpha_{\text{relax}})^2}}$$

which follows from the geometric definition of y_ζ given in Figure 4.2.

For temperature-constrained parameters $\beta \ge \frac{\pi}{\eta}$ corresponding to the innermost ellipse in Figure 4.6d, we have that $D(\zeta) = \emptyset$, $b^\star(\zeta) = [-1,1]$, $\alpha_{\min}(\zeta) = 0$ and thus

$$\alpha_{\text{diag}}(\zeta) = \alpha_{\text{anti}}(\zeta) = \alpha_{\max}(\zeta) \sim \beta^{-1}$$

where in the last step we again employed Lemma 4.7.8. It hence remains to analyze the asymptotics of $\alpha_{\mathrm{diag}}(\zeta)$ and $\alpha_{\mathrm{anti}}(\zeta)$, to which end we introduce

$$x^\star(\zeta) := \operatorname*{arg\,min}_{x \in \partial b^\star(\zeta)} \alpha_{b^\star(\zeta)}(x) = \operatorname*{arg\,max}_{x \in b^\star(\zeta)} \alpha_{[-1,1]}(x), \tag{4.29}$$

which is the point where $D(\zeta)$ and $E\big(-\alpha_{\min}(\zeta)\big)$ touch, see Figure 4.6c for an illustration of the analogous variable $\hat{x}^\star(s)$. We observe the following.

Lemma 4.7.9 *$x^\star(\zeta)$ is unique for $|\omega|$ small enough, and $\lim_{\omega\to 0} \operatorname{Re}\big(x^\star(\zeta)\big) = 0$.*

Proof. One verifies from the geometric interpretation of $x^\star(\zeta)$ in Figure 4.6c that $x^\star(\zeta)$ is unique and satisfies $\operatorname{Re}\big(x^\star(\zeta)\big) = 0$ if $\omega = 0$. The uniqueness and limit then follow from the continuity of $\alpha_{[-1,1]}(x)$ and $b^\star(\zeta)$. □

Lemma 4.7.10 *$\alpha_{\min}(\zeta) = -\alpha_{[-1,1]}\big(x^\star(\zeta)\big)$ and $\alpha_{\max}(\zeta) = \alpha_{[-1,1]}\big(x^\star(\zeta) - \omega - i\eta\big)$.*

Proof. The claim follows directly from (4.29). □

Lemma 4.7.11 *In the limit considered in Theorem 4.4.1 ($\beta \to \infty$ and $\omega, \eta \to 0$ with $|\omega| \lesssim \eta$), we have that*

$$\big|\operatorname{Im}\big(x^\star(\zeta)\big)\big| \sim \begin{cases} \eta^{1/2} & \text{if } \beta \le \frac{\pi}{y_\zeta}, \\ \beta^{-1} & \text{if } \beta \ge \frac{\pi}{y_\zeta}. \end{cases}$$

and thus $\eta = \mathcal{O}\big(|x^\star(\zeta)|^2\big)$

Proof. We conclude from Lemma 4.7.9 that for small ω, $x^\star(\zeta)$ is near the imaginary axis where $b^\star(\zeta) \ni x^\star(\zeta)$ satisfies

$$\operatorname{Im}\big(b^\star(\zeta)\big) = \big\{ \sinh\big(\alpha_{\max}(\zeta)\big)\, \sin(\theta) - \eta \mid \theta \approx \tfrac{3\pi}{2} \big\}.$$

The claim then follows from the asymptotics for $\alpha_{\max}(\zeta)$ which may be derived from Lemma 4.7.8. □

Using the above results and the shorthand notation $s = \omega + i\eta$, we get for the diagonal decay rate

$$\begin{aligned}
\alpha_{\mathrm{diag}}(\zeta) &= \tfrac{1}{2}\left(\alpha_{\max}(\zeta) + \alpha_{\min}(\zeta)\right) \\
&= \tfrac{1}{2}\Big(\alpha_{[-1,1]}\big(x^\star(\zeta) - s\big) - \alpha_{[-1,1]}\big(x^\star(\zeta)\big)\Big) \\
&= \frac{\partial \alpha_{[-1,1]}}{\partial \operatorname{Im}(x)}(0)\ \operatorname{Im}\big(x^\star(\zeta) - s - x^\star(\zeta)\big) + \ldots \\
&\qquad \operatorname{Re}\left(\frac{\partial^2}{\partial x^2} \log \phi_{[-1,1]}^{-1}(x)\Big|_{x=0} \Big(\big(x^\star(\zeta) - s\big)^2 - x^\star(\zeta)^2\Big)\right) + \mathcal{O}\big(|x^\star(\zeta)|^{3/2}\big) \\
&= \frac{\partial \alpha_{[-1,1]}}{\partial \operatorname{Im}(x)}(0)\, \eta + \mathcal{O}\big(|x^\star(\zeta)|^{3/2}\big) \\
&\sim \eta.
\end{aligned}$$

For the anti-diagonal decay rate $\alpha_{\mathrm{anti}}(\zeta)$, on the other hand, we repeat the above calculations with a negative sign for $\alpha_{\min}(\zeta)$, which means that the $x^\star(\zeta)$ in the linear term and the $x^\star(\zeta)^2$ in the quadratic term on the third line add up rather than cancel and thus

$$\alpha_{\mathrm{anti}}(\zeta) = \frac{\partial \alpha_{[-1,1]}}{\partial \operatorname{Im}(x)}(0)\ \operatorname{Im}\big(x^\star(\zeta)\big) + \mathcal{O}\big(|x^\star(\zeta)|^2\big) \sim \begin{cases} \eta^{1/2} & \text{if } \beta \le \frac{\pi}{y_\zeta}, \\ \beta^{-1} & \text{if } \beta \ge \frac{\pi}{y_\zeta}. \end{cases}$$

This completes the proof of Theorem 4.4.1.

4.8 Other Proofs

4.8.1 Proof of Theorem 4.4.2

Let us introduce

$$b_{k_1 k_2} := \exp\big(-\alpha_{\max}(\zeta)\, k_1 - \alpha_{\min}(\zeta)\, k_2\big)$$

with

$$\alpha_{\max}(\zeta) := \alpha_{\mathrm{diag}}(\zeta) + \alpha_{\mathrm{anti}}(\zeta), \qquad \alpha_{\min}(\zeta) := \alpha_{\mathrm{diag}}(\zeta) - \alpha_{\mathrm{anti}}(\zeta).$$

Using the triangle equality and $\|T_k\|_{[-1,1]} = 1$ and the bound (4.13), we obtain

$$\begin{aligned}
\|f - f_\tau\|_{[-1,1]^2} &\leq \sum_{(k_1,k_2)\in\mathbb{N}^2\setminus K(\tau)} |c_{k_1 k_2}| \\
&\leq 2\,C(\zeta) \sum_{(k_1,k_2)\in\mathbb{N}^2\setminus K(r)\wedge k_1\geq k_2} b_{k_1 k_2} \\
&= 2\,C(\zeta) \Bigg(\underbrace{\sum_{k_2=0}^{K_2(\tau)-1} \sum_{k_1=K_1(\tau,k_2)}^{\infty} b_{k_1 k_2}}_{A} + \underbrace{\sum_{k_2=K_2(\tau)}^{\infty} \sum_{k_1=k_2}^{\infty} b_{k_1 k_2}}_{B} \Bigg)
\end{aligned}$$

where

$$K_2(\tau) := \left\lceil \frac{-\log(\tau)}{2\,\alpha_{\mathrm{diag}}(\zeta)} \right\rceil, \qquad K_1(\tau, k_2) := \left\lceil -\frac{\log(\tau) + \alpha_{\min}(\zeta)\,k_2}{\alpha_{\max}(\zeta)} \right\rceil.$$

For the two terms A and B, we obtain using $\alpha_{\mathrm{diag}}(\zeta) = \mathcal{O}\big(\alpha_{\mathrm{anti}}(\zeta)\big)$ and hence $\alpha_{\max}(\zeta) = \Theta\big(\alpha_{\mathrm{anti}}(\zeta)\big)$,

$$\begin{aligned}
A &= \sum_{k_2=0}^{K_2(\tau)-1} \exp\big(-\alpha_{\min}(\zeta)\,k_2\big) \sum_{k_1=K_1(\tau,k_2)}^{\infty} \exp\big(-\alpha_{\max}(\zeta)\,k_1\big) \\
&\leq \sum_{k_2=0}^{K_2(\tau)-1} \exp\big(-\alpha_{\min}(\zeta)\,k_2\big) \frac{\tau\,\exp\big(\alpha_{\min}(\zeta)\,k_2\big)}{1-\exp\big(-\alpha_{\max}(\zeta)\big)} \\
&= \frac{K_2(\tau)}{1-\exp\big(-\alpha_{\max}(\zeta)\big)}\,\tau \\
&\lesssim \alpha_{\mathrm{diag}}(\zeta)^{-1}\,\alpha_{\mathrm{anti}}(\zeta)^{-1}\,\tau\,\log(\tau)
\end{aligned}$$

and

$$\begin{aligned}
B &= \sum_{k_2=K_2(\tau)}^{\infty} \exp\big(-\alpha_{\min}(\zeta)\,k_2\big) \sum_{k_1=k_2}^{\infty} \exp\big(-\alpha_{\max}(\zeta)\,k_1\big) \\
&= \sum_{k_2=K_2(\tau)}^{\infty} \exp\big(-\alpha_{\mathrm{diag}}(\zeta)\,k_2\big) \frac{1}{1-\exp\big(-\alpha_{\max}(\zeta)\big)} \\
&\leq \frac{\tau}{1-\exp\big(-\alpha_{\mathrm{diag}}(\zeta)\big)}\,\frac{1}{1-\exp\big(-\alpha_{\max}(\zeta)\big)} \\
&\lesssim \alpha_{\mathrm{diag}}(\zeta)^{-1}\,\alpha_{\mathrm{anti}}(\zeta)^{-1}\,\tau.
\end{aligned}$$

This completes the proof of Theorem 4.4.2.

4.8.2 Proof of Theorem 4.5.1

According to Riemann's removable singularity theorem in higher dimensions (see e.g. [Sch05, Thm. 4.2.1]), the function

$$R(E_1, E_2) = \big(E_1 - E_2 + \omega + i\eta\big)\, F_\zeta(E_1, E_2) - \frac{1}{\beta}\, \frac{1}{(E_1 - z)\,(E_2 - z)} \tag{4.30}$$

with $z := \frac{\pi i}{\beta}$ can be analytically continued to

$$\mathcal{S}_z := \Big(\{z\} \times \big(\mathbb{C} \setminus \mathcal{S}_{\beta,E_F}\big)\Big) \cup \Big(\big(\mathbb{C} \setminus \mathcal{S}_{\beta,E_F}\big) \times \{z\}\Big)$$

if $R(E_1, E_2)$ is bounded on this set, or equivalently if

$$\lim_{E_1 \to z} (E_1 - z)\, R(E_1, E_2) = 0 \tag{4.31}$$

for some arbitrary $E_2 \in \mathbb{C} \setminus \mathcal{S}_{\beta,E_F}$ and likewise with the roles of E_1 and E_2 interchanged. In order to verify (4.31), we compute

$$\begin{aligned} \lim_{E_1 \to z} (E_1 - z)\, f_{\text{temp}}(E_1, E_2) &= \lim_{E_1 \to z} (E_1 - z)\, \frac{f_{\beta,E_F}(E_1) - f_{\beta,E_F}(E_2)}{E_1 - E_2} \\ &= \frac{1}{z - E_2}\, \lim_{E_1 \to z} \frac{E_1 - z}{1 + \exp\big(\beta\,(E_1 - E_F)\big)} \\ &= \frac{1}{\beta}\, \frac{1}{E_2 - z} \end{aligned} \tag{4.32}$$

where on the last line we used L'Hôpital's rule to determine the limit. It follows from (4.32) that for $E_1 \to z$, the first and second term in (4.30) cancel and hence (4.31) holds. The transposed version of (4.31) follows from the symmetry of (4.30); thus we conclude that $R(E_1, E_2)$ can indeed be analytically continued to $\mathcal{S}_z$. Theorem 4.5.1 then follows by rewriting (4.15) in the form (4.30) and applying the above argument to each of the terms in the sum over Z_k.

4.8.3 Proof of Theorem 4.1.4

It follows from (4.8) that the local conductivity σ_{loc} can be written in the form

$$\sigma_{\text{loc}} = \sum_{k_1,k_2=0}^{\infty} c_{k_1 k_2}\, \Big(T_{k_1}(H_{\text{loc}})\, M_a^{\text{loc}}\, T_{k_2}(H_{\text{loc}})\, M_b^{\text{loc}}\Big)_{0,\alpha_\ell;0,\alpha_\ell},$$

and we conclude from an argument analogous to the one given in Lemma 3.2.3 that the weights

$$w_{k_1k_2} := \Big(T_{k_1}(H_{\rm loc})\, M_a^{\rm loc}\, T_{k_2}(H_{\rm loc})\, M_b^{\rm loc}\Big)_{0,\alpha_\ell;0,\alpha_\ell}$$

become independent of the buffer radius r for $r \geq C\,(k_1 + k_2)$ with a constant C depending on the sparsity structure of $M_a^{\rm loc}$ and $H_{\rm loc}$. We have already observed in Lemma 4.2.1 that $w_{k_1k_2}$ is bounded for all r and k_1, k_2; hence the difference in Theorem 4.1.4 may be estimated by

$$\big|\sigma^r_{\ell,\alpha_\ell}\big(A_{T(\ell)}\,u\big) - \sigma_{\ell,\alpha_\ell}\big(A_{T(\ell)}\,u\big)\big| \lesssim \sum_{k_1+k_2 \geq \frac{r}{C}} |c_{k_1k_2}| \lesssim_\varepsilon \exp\big(-\tfrac{1}{C}\,\alpha_{\rm diag}(\zeta)\,r\big)$$

as claimed.

4.9 Conclusion

We developed an algorithm for conductivity calculations on incommensurate bilayers based on the ergodicity property (4.3) and a combination of ideas from the domain decomposition and function approximation approaches. The proposed method requires a polynomial or rational approximation p to the conductivity function (4.2) and its cost is minimised if we choose p to have as few terms as possible while meeting the accuracy requirement. Our main contribution in this chapter has been the analysis of this approximation problem, which showed that the proposed local conductivity algorithm scales with only $\mathcal{O}\big((\eta^{-3/2} + \beta^{1/2}\,\eta^{-5/4} + \beta^{2/3}\,\eta^{-1})\,m\big)$ (cf. Theorem 4.5.2) rather than $\mathcal{O}\big((\eta^{-2} + \beta^2)\,m\big)$ as one might have expected based on one-dimensional arguments (m denotes the matrix size of the local Hamiltonian). Furthermore, our analysis illustrates several surprising and (to the best of the author's knowledge) little known features of two-dimensional approximation theory which we would like to emphasise in the following.

Degree vs. number of terms. The performance of a polynomial approximation scheme $p_n(x) \approx f(x)$ is usually measured in terms of how the error $e_n = \|p_n - f\|_{[-1,1]}$ decays as a function of the polynomial degree n. This is well justified in one dimension since the cost of evaluating the powers x^k (the degree) is proportional to the cost of summing them (the number of terms) except in the rare case when the approximand is of the form $p_n(x) = \sum_{k=0}^n c_k\, x^k$ with a sparse coefficient vector c_k. In two dimensions, one might expect that the relationship between the degree n and the number of terms $n_{\rm terms}$ would consistently be $n_{\rm terms} = \mathcal{O}\big(n^2\big)$, but we have seen with the example of the relaxation-constrained conductivity function that this may be far from the truth and in fact the relationship could be anywhere between $n_{\text{terms}} = \mathcal{O}(n)$ and $n_{\text{terms}} = \mathcal{O}(n^2)$. This suggests that in dimensions $d > 1$, the performance of a polynomial approximation scheme is more reasonably measured in terms of error as a function of the number of terms, though even this metric has its limitations as pointed out in Subsection 4.6.2.

Representation. An immediate corollary of the previous point is the observation that in dimensions $d > 1$, the representation of a polynomial impacts the efficiency with which it can be evaluated in a non-trivial way. More precisely, while in one dimension all commonly used representations of a polynomial of degree n allow for evaluation in $\mathcal{O}(n)$ floating-point operations, we have seen that if we approximate the conductivity function in the Chebyshev basis, the resulting coefficients exhibit a form of sparsity which greatly reduces the cost of evaluation but which would have been absent if we had represented our approximation e.g. in barycentric form, because then the expansion coefficients are given by point values of F_ζ which exhibit no sparsity.

Stability vs. efficiency. The most efficient way to exploit the rational approximation ideas presented in Section 4.5 would be to merge all of the troublesome low-temperature poles $Z := \mathcal{S}_{\beta,E_F} \cap E(\alpha_{\text{relax}})$ into a single factor $q(E) = \prod_{z \in Z}(E - z)^{-1}$ and approximate $p(E_1, E_2) \approx F_\zeta(E_1, E_2)\, q(E_1)^{-1}\, q(E_2)^{-1}$ using the truncated Chebyshev series from Section 4.4. However, such a scheme would be highly unstable as pointed out in Remark 4.5.3, and this forced us to compromise between evaluation efficiency and stability in Algorithm 4.2. It remains unclear whether such compromises are an inherent difficulty in high-dimensional rational approximation or whether they can be avoided by some other approximation scheme.

Radius of analyticity and rate of decay. Our theory regarding the decay of the Chebyshev coefficients in two dimensions in Section 4.7 closely mirrors the corresponding one-dimensional theory which is known to predict the exact asymptotic decay rates. We were therefore surprised to find that this theory failed to predict the non-convex shape of the Chebyshev coefficients of the far-relaxation-constrained conductivity function in Figure 4.3a, but it turns out that this is an inherent feature of higher-dimensional Chebyshev series: unlike in one dimension, the asymptotic behaviour of the Chebyshev coefficients is not fully determined by the domain of analyticity of the represented function as demonstrated in the following example.

Example 4.9.1 The two functions $f_+(x_1, x_2) := g_1(x_1) + g_2(x_2)$ and $f_\times(x_1, x_2) := g_1(x_2)\, g_2(x_2)$ have the same domain of analyticity, but the Chebyshev coefficients $c[f_+]_{k_1 k_2} = c[g_1]_{k_1}\, \delta_{k_2 0} + \delta_{k_1 0}\, c[g_2]_{k_2}$ of f_+ are zero if $k_1, k_2 > 0$, while the coefficients $c[f_\times]_{k_1 k_2} = c[g_1]_{k_1}\, c[g_2]_{k_2}$ of $f_\times$ are generally nonzero for any $k_1, k_2 \in \mathbb{N}$.

We would like to conclude this chapter by discussing two more pairs of functions $f(x_1, x_2)$ and their associated Chebyshev coefficients $c_{k_1 k_2}$. This serves on the one hand to demonstrate that the theory developed in Section 4.7 for the conductivity function easily extends to other functions, and on the other hand to highlight the subtlety and richness of polynomial approximation in higher dimensions. In both examples, we will follow the notation from Section 4.7 and in particular Subsection 4.7.3.

Example 4.9.2 Consider the function

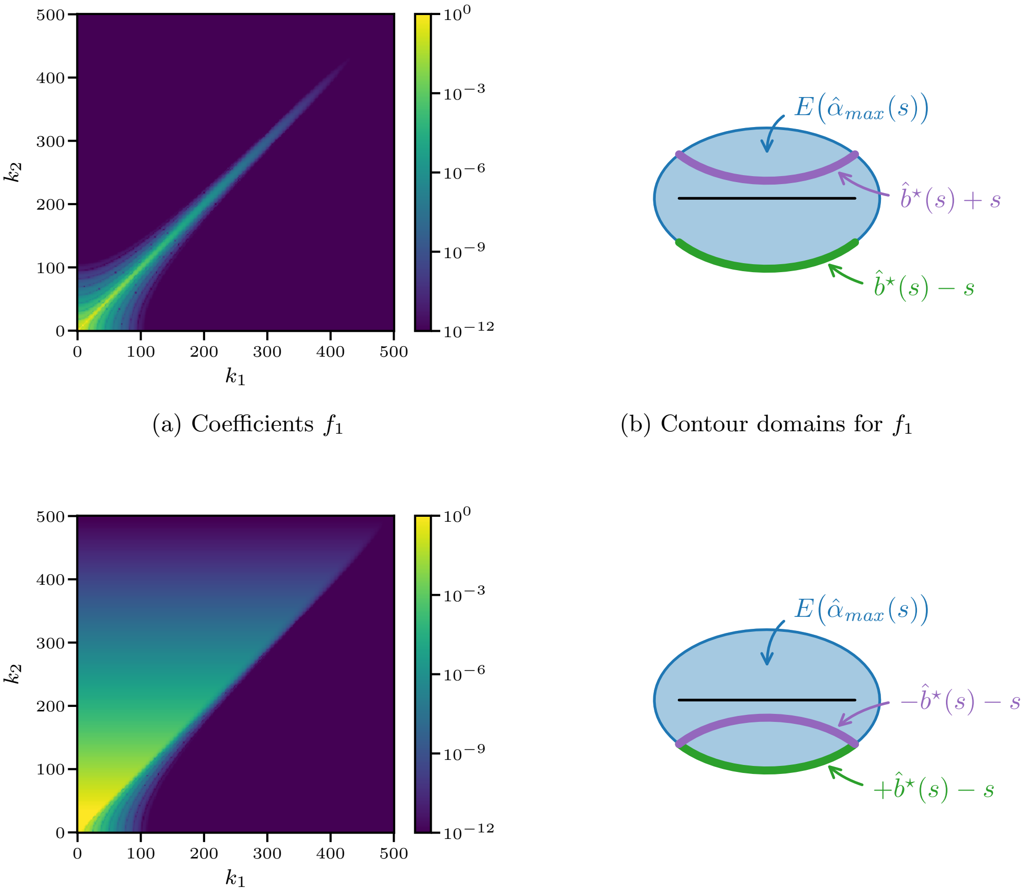
$$f_1(x_1, x_2) := \frac{1}{(x_1 - x_2)^2 - s^2}$$

for some s on the positive imaginary axis. The singularities in x_1 for a given x_2 are $x_1 = x_2 \pm s$ which suggests that there should be no concentration along the diagonal since if we try to replace $\Omega_2 = [-1, 1]$ with $\Omega_2 = \hat{b}^\star(s)$ to make space for a larger ellipse $\Omega_1 = E\big(\hat{\alpha}_{\max}(s)\big)$ as in Figure 4.6b, then the singularities $\hat{b}^\star(s) + s$ in the upper half-plane will penetrate this ellipse $\Omega_1 = E\big(\hat{\alpha}_{\max}(s)\big)$, see Figure 4.7b. However, we do observe concentration along the diagonal in Figure 4.7a, and the modification required to reconcile this observation with the above argument is to note that we may choose the contour integral domain Ω_2 in x_2 depending on our position $x_1 \in \partial\Omega_1$ in the contour integral of x_1. This allows us to choose $\Omega_2(x_1) = \hat{b}^\star(s)$ if x_1 is in the lower half-plane and $\Omega_2(x_1) = -\hat{b}^\star(s)$ if x_1 is in the upper half-plane, which in turn enables us to choose $\Omega_1 = E\big(\hat{\alpha}_{\max}(s)\big)$.

Example 4.9.3 Our second example is the function

$$f_2(x_1, x_2) := \frac{1}{(x_1 + s)^2 - x_2^2}$$

where s is again some arbitrary point on the positive imaginary axis. As shown in Figure 4.7c, the Chebyshev coefficients of this function do not concentrate along the diagonal in the region $k_1 > k_2$ which may be explained as follows. For a fixed x_2, the singularities in x_1 are given by $x_1 = \pm x_2 - s$; hence if we try to move the point $x_2 = 0$ in the direction of the negative imaginary axis to make space for a larger ellipse in x_1, then the other singularity will move in the direction of the positive imaginary axis which decreases the ellipse of analyticity in x_1. It follows that $\Omega_2 = [-1, 1]$ is in fact the best possible choice for maximising Ω_1, which explains Figure 4.7c.



(a) Coefficients f_1

(b) Contour domains for f_1

(c) Coefficients f_2

(d) Contour domains for f_2

Figure 4.7: (a,c) Normalized Chebyshev coefficients $\hat{c}_{k_1 k_2} := |c_{k_1 k_2}|/|c_{00}|$ of the functions f_1, f_2 from Examples 4.9.2 and 4.9.3, respectively, with $s = 0.06i$. Only the coefficients with even indices k_1, k_2 are shown to hide the zero coefficients introduced by the symmetries of the function. (b,d) Some of the contour domains mentioned in Examples 4.9.2 and 4.9.3.

Appendix

A.1 Hardware and Software for Numerical Experiments

All numerical experiments in this thesis have been performed on a single core of an Intel Core i7-8550 CPU (1.8 GHz base frequency, 4 GHz turbo boost) using the Julia programming language [BEKS17]. Plots were created using Matplotlib [Hun07] and TikZ [Tik]. Several Julia packages developed as part of this thesis are available online at `github.com/ettersi`.

A.2 Asymptotic Relations

Given two functions $f, g : \mathbb{R} \to \mathbb{R}$ and a limit point $x_0 \in \mathbb{R}$, we write "$f(x) \lesssim g(x)$ for $x \to x_0$" if there exists neighbourhood N of x_0 and a constant $C > 0$ such that $f(x) \leq C\, g(x)$ for all $x \in N$. Furthermore, we write $f(x) \gtrsim g(x)$ if $g(x) \lesssim f(x)$, and we write $f(x) \sim g(x)$ if both $f(x) \lesssim g(x)$ and $f(x) \gtrsim g(x)$.

Bibliography

[ADD96] P. R. Amestoy, T. A. Davis, and I. S. Duff, *An approximate minimum degree ordering algorithm*, SIAM Journal on Matrix Analysis and Applications, 17 (1996), pp. 886–905, doi:10.1137/S0895479894278952.

[ADD04] P. R. Amestoy, T. A. Davis, and I. S. Duff, *Algorithm 837: AMD, an approximate minimum degree ordering algorithm*, ACM Transactions on Mathematical Software, 30 (2004), pp. 381–388, doi:10.1145/1024074.1024081.

[AG18] N. H. Asmar and L. Grafakos, *Complex Analysis with Applications*, Undergraduate Texts in Mathematics, Springer, Cham, 2018, doi:10.1007/978-3-319-94063-2.

[ARC] *Application usage over past month on ARCHER*, `http://www.archer.ac.uk/status/codes/`. Last accessed on 5 May 2019.

[BBR13] M. Benzi, P. Boito, and N. Razouk, *Decay properties of spectral projectors with applications to electronic structure*, SIAM Review, 55 (2013), pp. 3–64, doi:10.1137/100814019.

[BEKS17] J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, *Julia: A fresh approach to numerical computing*, SIAM Review, 59 (2017), pp. 65–98, doi:10.1137/141000671.

[BM48] S. Bochner and W. T. Martin, *Several Complex Variables*, Princeton University Press, 1948.

[BM10] D. R. Bowler and T. Miyazaki, *Calculations for millions of atoms with density functional theory: linear scaling shows its potential*, Journal of Physics: Condensed Matter, 22 (2010), doi:10.1088/0953-8984/22/7/074207.

[BM12] D. R. Bowler and T. Miyazaki, *O(N) methods in electronic structure calculations.*, Reports on Progress in Physics, 75 (2012), doi:10.1088/0034-4885/75/3/036503.

[Boy09] J. P. Boyd, *Large-degree asymptotics and exponential asymptotics for Fourier, Chebyshev and Hermite coefficients and Fourier transforms*, Journal of Engineering Mathematics, 63 (2009), pp. 355–399, doi:10.1007/s10665-008-9241-3.

[Bra86] D. Braess, *Nonlinear Approximation Theory*, vol. 7 of Springer Series in Computational Mathematics, Springer, Berlin, Heidelberg, 1986, doi:10.1007/978-3-642-61609-9.

[CFF$^+$18] Y. Cao, V. Fatemi, S. Fang, K. Watanabe, T. Taniguchi, E. Kaxiras, and P. Jarillo-Herrero, *Unconventional superconductivity in magic-angle graphene superlattices*, Nature, 556 (2018), pp. 43–50, doi:10.1038/nature26160.

[CO16] H. Chen and C. Ortner, *QM/MM methods for crystalline defects. Part 1: locality of the tight binding model*, Multiscale Modeling & Simulation, 14 (2016), pp. 232–264, doi:10.1137/15M1022628.

[Coa18] M. Coates, *Polynomial and Rational Approximation of the Fermi-Dirac Function*, master's thesis, University of Warwick, 2018.

[CP15] E. Chow and A. Patel, *Fine-grained parallel incomplete LU factorization*, SIAM Journal on Scientific Computing, 37 (2015), pp. C169–C193, doi:10.1137/140968896.

[CSC17] *Annual Report of the Swiss National Supercomputing Centre*, 2017, `https://www.cscs.ch/publications/annual-reports/cscs-annual-report-2017/`.

[Dav06] T. A. Davis, *Direct Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, 2006, doi:10.1137/1.9780898718881.

[DC11] X. Dong and G. Cooperman, *A bit-compatible parallelization for ILU(k) preconditioning*, in European Conference on Parallel Processing, 2011, pp. 66–77, doi:10.1007/978-3-642-23397-5_8.

[Dev06] L. Devroye, *Nonuniform random variate generation*, in Handbooks in Operations Research and Management Science, vol. 13, 2006, ch. 4, pp. 83–121, doi:10.1016/S0927-0507(06)13004-2.

[DMS84] S. Demko, W. F. Moss, and P. W. Smith, *Decay rates for inverses of band matrices*, Mathematics of Computation, 43 (1984), pp. 491–499, doi:10.1090/S0025-5718-1984-0758197-9.

[Dri96] T. A. Driscoll, *Algorithm 756: A MATLAB toolbox for Schwarz-Christoffel mapping*, ACM Transactions on Mathematical Software, 22 (1996), pp. 168–186, doi:10.1145/229473.229475.

[DT02] T. A. Driscoll and L. N. Trefethen, *Schwarz-Christoffel Mapping*, Cambridge University Press, 2002, doi:10.1017/CBO9780511546808.

[EMLO19] S. Etter, D. Massatt, M. Luskin, and C. Ortner, *Modeling and computation of Kubo conductivity for 2D incommensurate bilayers*, submitted, (2019), arXiv:1907.01314.

[ET75] A. M. Erisman and W. F. Tinney, *On computing certain elements of the inverse of a sparse matrix*, Communications of the ACM, 18 (1975), pp. 177–179, doi:10.1145/360680.360704.

[ET99] M. Embree and L. N. Trefethen, *Green's functions for multiply connected domains via conformal mapping*, SIAM Review, 41 (1999), pp. 745–761, doi:10.1137/S0036144598349277.

[GC94] S. Goedecker and L. Colombo, *Efficient linear scaling algorithm for tight-binding molecular dynamics*, Physical Review Letters, 73 (1994), pp. 122–125, doi:10.1103/PhysRevLett.73.122.

[Geo73] A. George, *Nested dissection of a regular finite element mesh*, SIAM Journal on Numerical Analysis, 10 (1973), pp. 345–363, doi:10.1137/0710032.

[GG13] A. K. Geim and I. V. Grigorieva, *Van der Waals heterostructures*, Nature, 499 (2013), pp. 419–425, doi:10.1038/nature12385.

[Gil88] J. R. Gilbert, *Some nested dissection order is nearly optimal*, Information Processing Letters, 26 (1988), pp. 325–328, doi:10.1016/0020-0190(88)90191-3.

[Goe99] S. Goedecker, *Linear scaling electronic structure methods*, Reviews of Modern Physics, 71 (1999), pp. 1085–1123, doi:10.1103/RevModPhys.71.1085.

[GT95] S. Goedecker and M. Teter, *Tight-binding electronic-structure calculations and tight-binding molecular dynamics with localized orbitals*, Physical Review B, 51 (1995), pp. 9455–9464, doi:10.1103/PhysRevB.51.9455.

[GV96] G. H. Golub and C. F. Van Loan, *Matrix Computations*, Johns Hopkins University Press, Baltimore, third ed., 1996.

[HHT08] N. Hale, N. J. Higham, and L. N. Trefethen, *Computing A^α, $\log(A)$, and related matrix functions by contour integrals*, SIAM Journal on Numerical Analysis, 46 (2008), pp. 2505–2523, doi:10.1137/070700607.

[Hin17] N. D. Hine, *Linear-scaling density functional theory using the projector augmented wave method*, Journal of Physics Condensed Matter, 29 (2017), doi:10.1088/0953-8984/29/2/024001.

[HP01] D. Hysom and A. Pothen, *A scalable parallel algorithm for incomplete factor preconditioning*, SIAM Journal on Scientific Computing, 22 (2001), pp. 2194–2215, doi:10.1137/S1064827500376193.

[Hun07] J. D. Hunter, *Matplotlib: A 2D graphics environment*, Computing in Science & Engineering, 9 (2007), pp. 90–95, doi:10.1109/MCSE.2007.55.

[JLY16] M. Jacquelin, L. Lin, and C. Yang, *PSelInv – A distributed memory parallel algorithm for selected inversion*, ACM Transactions on Mathematical Software, 43 (2016), doi:10.1145/2786977.

[Kax03] E. Kaxiras, *Atomic and Electronic Structure of Solids*, Cambridge University Press, 2003, doi:10.1017/CBO9780511755545.

[KK97] G. Karypis and V. Kumar, *Parallel threshold-based ILU factorization*, in Proceedings of the 1997 ACM/IEEE conference on Supercomputing, 1997, doi:10.1145/509593.509621.

[Koh96] W. Kohn, *Density functional and density matrix method scaling linearly with the number of atoms*, Physical Review Letters, 76 (1996), pp. 3168–3171, doi:10.1103/PhysRevLett.76.3168.

[Kra01] S. G. Krantz, *Function Theory of Several Complex Variables*, American Mathematical Society, 2001.

[LCYH13] L. Lin, M. Chen, C. Yang, and L. He, *Accelerating atomic orbital-based electronic structure calculation via pole expansion and selected inversion*, Journal of Physics: Condensed Matter, 25 (2013), doi:10.1088/0953-8984/25/29/295501.

[LGHY14] L. Lin, A. García, G. Huhs, and C. Yang, *SIESTA-PEXSI: Massively parallel method for efficient and accurate ab initio materials simulation without matrix diagonalization*, Journal of Physics: Condensed Matter, 26 (2014), doi:10.1088/0953-8984/26/30/305503.

[LLY+09] L. Lin, J. Lu, L. Ying, R. Car, and W. E, *Fast algorithm for extracting the diagonal of the inverse matrix with application to the electronic structure analysis of metallic systems*, Communications in Mathematical Sciences, 7 (2009), pp. 755–777, doi:10.4310/CMS.2009.v7.n3.a12.

[LLYE09] L. Lin, J. Lu, L. Ying, and W. E, *Pole-based approximation of the Fermi-Dirac function*, Chinese Annals of Mathematics, Series B, 30 (2009), pp. 729–742, doi:10.1007/s11401-009-0201-7.

[LYM+11] L. Lin, C. Yang, J. C. Meza, J. Lu, L. Ying, and W. E, *SelInv – An algorithm for selected inversion of a sparse symmetric matrix*, ACM Transactions on Mathematical Software, 37 (2011), doi:10.1145/1916461.1916464.

[MLO17] D. Massatt, M. Luskin, and C. Ortner, *Electronic density of states for incommensurate layers*, Multiscale Modeling & Simulation, 15 (2017), pp. 476–499, doi:10.1137/16M1088363.

[Mou16] J. E. Moussa, *Minimax rational approximation of the Fermi-Dirac distribution*, The Journal of Chemical Physics, 145 (2016), doi:10.1063/1.4965886.

[MRG+15] S. Mohr, L. E. Ratcliff, L. Genovese, D. Caliste, P. Boulanger, S. Goedecker, and T. Deutsch, *Accurate and efficient linear scaling DFT calculations with universal applicability*, Physical Chemistry Chemical Physics, 17 (2015), pp. 31335–32058, doi:10.1039/c5cp00437c.

[Mul66] R. S. Mulliken, *Spectroscopy, molecular orbitals, and chemical bonding*, 1966, `https://www.nobelprize.org/prizes/chemistry/1966/mulliken/`.

[NF16] Y. Nakatsukasa and R. W. Freund, *Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: the power of Zolotarev's functions*, SIAM Review, 58 (2016), pp. 461–493, doi:10.1137/140990334.

[Olv11] S. Olver, *Computation of equilibrium measures*, Journal of Approximation Theory, 163 (2011), pp. 1185–1207, doi:10.1016/j.jat.2011.03.010.

[OTBM16] T. Otsuka, M. Taiji, D. R. Bowler, and T. Miyazaki, *Linear-scaling first-principles molecular dynamics of complex biological systems with the CONQUEST code*, Japanese Journal of Applied Physics, 55 (2016), doi:10.7567/JJAP.55.1102B1.

[Ran95] T. Ransford, *Potential Theory in the Complex Plane*, Cambridge University Press, 1995.

[RT78] D. J. Rose and R. E. Tarjan, *Algorithmic aspects of vertex elimination on directed graphs*, SIAM Journal on Applied Mathematics, 34 (1978), pp. 176–197, doi:10.1137/0134014.

[Saa03] Y. Saad, *Iterative Methods for Sparse Linear Systems*, Society for Industrial and Applied Mathematics, second ed., 2003, doi:10.1137/1.9780898718003.

[Saf10] E. B. Saff, *Logarithmic potential theory with applications to approximation theory*, Surveys in Approximation Theory, 5 (2010), pp. 165–200, https://www.emis.de/journals/SAT/papers/14/.

[SAG$^+$02] J. M. Soler, E. Artacho, J. D. Gale, A. García, J. Junquera, P. Ordejón, and D. Sánchez-Portal, *The SIESTA method for ab initio order-N materials simulation*, Journal of Physics: Condensed Matter, 14 (2002), pp. 2745–2779, doi:10.1088/0953-8984/14/11/302.

[Sch05] V. Scheidemann, *Introduction to Complex Analysis in Several Variables*, Birkhäuser, Basel, 2005.

[SCS10] Y. Saad, J. R. Chelikowsky, and S. M. Shontz, *Numerical methods for electronic structure calculations of materials*, SIAM Review, 52 (2010), pp. 3–54, doi:10.1137/060651653.

[SHMP05] C.-K. Skylaris, P. D. Haynes, A. A. Mostofi, and M. C. Payne, *Introducing ONETEP: Linear-scaling density functional simulations on parallel computers*, The Journal of Chemical Physics, 122 (2005), doi:10.1063/1.1839852.

[SRVK96] R. N. Silver, H. Roeder, A. F. Voter, and J. D. Kress, *Kernel polynomial approximations for densities of states and spectral functions*, Journal of Computational Physics, 124 (1996), pp. 115–130, doi:10.1006/jcph.1996.0048.

[SSW01] J. Shen, G. Strang, and A. J. Wathen, *The potential theory of several intervals and its applications*, Applied Mathematics and Optimization, 44 (2001), pp. 67–85, doi:10.1007/s00245-001-0011-0.

[ST97] E. B. Saff and V. Totik, *Logarithmic Potentials with External Fields*, vol. 316 of Grundlehren der mathematischen Wissenschaften, Springer, Berlin, Heidelberg, 1997, doi:10.1007/978-3-662-03329-6.

[SZW03] C. Shen, J. Zhan, and K. Wang, *Parallel multilevel block ILU preconditioning techniques for large sparse linear systems*, in Proceedings International Parallel and Distributed Processing Symposium, 2003, doi:10.1109/IPDPS.2003.1213182.

[TFC73] K. Takahashi, J. Fagan, and M.-S. Chin, *Formation of a sparse bus impedence matrix and its application to short circuit study*, in 8th PICA Conference Proceedings, 1973.

[Tik] *TikZ*, `https://github.com/pgf-tikz/pgf`.

[Tre13] L. N. Trefethen, *Approximation Theory and Approximation Practice*, Society for Industrial and Applied Mathematics, 2013.

[Tre17] L. N. Trefethen, *Multivariate polynomial approximation in the hypercube*, Proceedings of the American Mathematical Society, 145 (2017), pp. 4837–4844, doi:10.1090/proc/13623.

[TRM] *Transpose of rational matrix is also rational*, `https://math.stackexchange.com/q/3192042`. (18 April 2019).

[TW14] L. N. Trefethen and J. A. C. Weideman, *The exponentially convergent trapezoidal rule*, SIAM Review, 56 (2014), pp. 385–458, doi:10.1137/130932132.

[VKM+05] J. VandeVondele, M. Krack, F. Mohamed, M. Parrinello, T. Chassaing, and J. Hutter, *QUICKSTEP: Fast and accurate density functional calculations using a mixed Gaussian and plane waves approach*, Computer Physics Communications, 167 (2005), pp. 103–128, doi:10.1016/j.cpc.2004.12.014.

[VKS96] A. F. Voter, J. D. Kress, and R. N. Silver, *Linear-scaling tight binding from a truncated-moment approach*, Physical Review B, 53 (1996), pp. 12733–12741, doi:10.1103/PhysRevB.53.12733.

[Wal56] J. L. Walsh, *Interpolation and Approximation by Rational Functions in the Complex Domain*, American Mathematical Society, second ed., 1956.

[Yan81] M. Yannakakis, *Computing the minimum fill-in is NP-complete*, SIAM Journal on Algebraic Discrete Methods, 2 (1981), pp. 77–79, doi:10.1137/0602010.

[YCG+18] V. W.-z. Yu, F. Corsetti, A. García, W. P. Huhn, M. Jacquelin, W. Jia, B. Lange, L. Lin, J. Lu, W. Mi, A. Seifitokaldani, Á. Vázquez-Mayagoitia, C. Yang, H. Yang, and V. Blum, *ELSI: A unified software interface for Kohn-Sham electronic structure solvers*, Computer Physics Communications, 222 (2018), pp. 267–285, doi:10.1016/j.cpc.2017.09.007.

[Zol77] E. I. Zolotarev, *Application of elliptic functions to questions of functions deviating least and most from zero*, Zapiskah Rossijskoi Akad. Nauk., (1877).