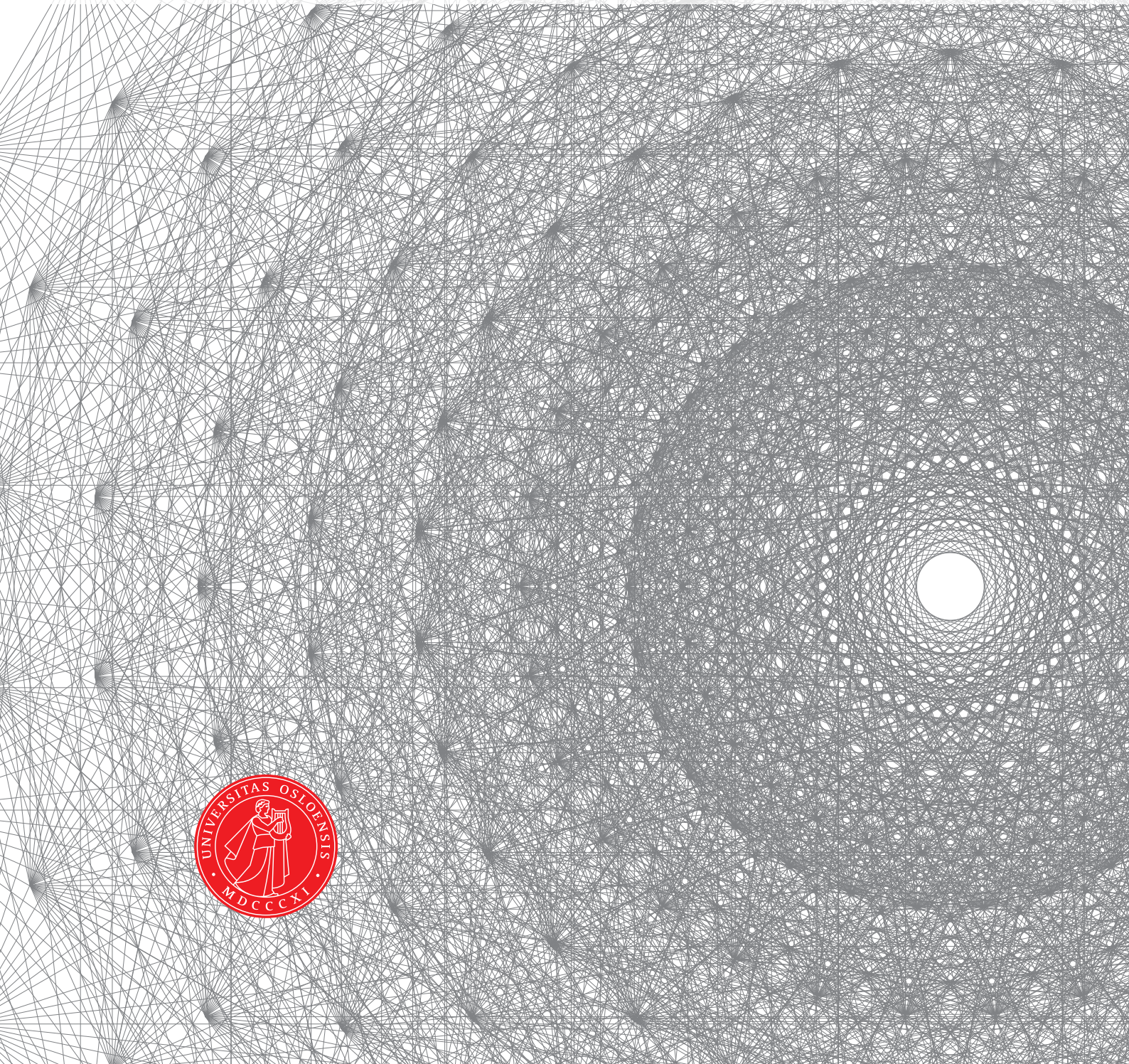
UiO : Department of Mathematics
University of Oslo

Computing Spectra with Error Control

With applications to Dirac operators

Emil Haugen

Master's Thesis, Spring 2022



This master's thesis is submitted under the master's programme *Mathematics*, with programme option *Mathematics*, at the Department of Mathematics, University of Oslo. The scope of the thesis is 60 credits.

The front page depicts a section of the root system of the exceptional Lie group E_8, projected into the plane. Lie groups were invented by the Norwegian mathematician Sophus Lie (1842–1899) to express symmetries in differential equations and today they play a central role in various parts of mathematics.

Abstract

This thesis concerns the problem of computing spectra of unbounded linear operators acting on infinite dimensional Hilbert spaces. Combining ideas from operator theory, numerical analysis and computability theory, we show how to construct algorithms which, given an operator, converge to its spectrum in a suitable topology. The key feature of these algorithms is that they converge to the true spectrum "from below" in a well defined sense, providing a so-called Σ_1-algorithm with rigorous control on the approximation error. We then apply this theory to self-adjoint Dirac operators from quantum mechanics. Specifically, we consider three-dimensional Dirac operators with bounded potentials and two-dimensional Dirac operators with infinite mass boundary conditions. In both cases, we manage to obtain a Σ_1-classification, hence proposing a numerical algorithm to compute the spectrum where analytical methods are not available.

Acknowledgements

First of all, I would like to thank my supervisor Anders Hansen for introducing me to computational spectral theory and his suggestion to look for applications to Dirac operators. Learning new theory and finding areas of application has been very educational. I am also grateful for Matthew Colbrook's patient and helpful feedback both in the reading of the background material and in discussion of applications. Thanks to Fabio Pizzichillo and Biagio Cassano for graciously answering all of my questions as I tried to wrap my head around their papers which formed the basis for key parts of this thesis. Thanks to supervisor Øyvind Ryan for valuable feedback on the writing process.

I am very happy to have spent my years as a student at the University of Oslo, in large part due to my friends and fellow students who have made it a joy to be a student in general, and a mathematics student in particular. Last but certainly not least, thanks to Camilla for sharing her knowledge on academic writing and being a kind, patient supporter.

Contents

CHAPTER 1

Introduction and Outline

Operators acting on infinite dimensional Hilbert spaces play a fundamental role in all of mathematical analysis and modern physics. For example, the basic postulates of quantum mechanics say that the possible values for any observable quantity (e.g. position, momentum, energy) of an isolated physical system (e.g. an electron in orbit around a nucleus) is given by the spectrum of a self-adjoint operator [10]. Thus a fundamental question in mathematical analysis is: Given a linear operator, how do we find its spectrum?

There is a vast literature on the problem of computing and using eigenvalues of finite dimensional matrices. The key term here is *finite dimensional*, because even though the theory of *infinite dimensional* spaces predates the study of finite matrices, the problem of finding algorithms that approximate spectra in infinite dimension has remained unsolved. As pointed out by W. Arveson [2] in 1993: "*Unfortunately, there is a dearth of literature on this basic problem and, so far as we have been able to tell, there are no proven techniques*".

In any approximation it is highly desirable to have control over how large the error may be. Hence, the topic of this thesis is algorithms which, given an operator, outputs sets that approximate its spectrum, while providing rigorous error estimates. The computational framework we use to study this problem is the Solvability Complexity Index (SCI) hierarchy, first introduced in [16]. Using this framework, the problem of approximating spectra of Schrödinger operators from point samples of the potential was considered in [5], while [8] studied general differential operators. Following the same overall approach, we give similar results for two different classes of Dirac operators, producing algorithms which compute the spectrum "from below".

The structure of the thesis is as follows: Chapter 2 contains preliminaries regarding unbounded operators. Chapter 3 presents the theory of computation and relevant algorithms. Chapter 4 gathers some necessary results on numerical integration. In Chapter 5 we formulate a setup for Dirac operators analogous to [5, Theorem 8.3] on Schrödinger operators, and give a corresponding result. Chapter 6 concerns two dimensional Dirac operators with boundary conditions as recently studied in [7]. Specifically, we show how the algorithms from Chapter 3 can be used to solve the problem of finding points in the discrete spectrum, which is left open by [7]. Thus the main contribution of this thesis is to show how the spectrum of these Dirac operators can be rigorously approximated using the methods described in [5, 8]. Concrete algorithms and pseudocode are given in Appendix A. A list of symbols is included at the end, while further notation is introduced throughout the thesis.

CHAPTER 2

Linear Operators in Hilbert Space

This chapter establishes the basic theory of linear operators on Hilbert spaces that will be needed. We assume that the basic theory of functional analysis, measure spaces and linear operators is familiar, with reference to e.g. [14, 18, 24]. The main objects of study in this thesis are operators which are *unbounded*. Though unbounded operators such as differentiation and multiplication operators are ubiquitous in applications, they are not typically covered in the standard sequence of analysis courses. Therefore, we devote some space and time to them in this chapter, basing the exposition on [24] and [26]. Section 2.1 provides basic definitions and results concerning unbounded operators and Section 2.2 discusses their spectra. Section 2.3 recounts some facts about tensor products in Hilbert spaces that will be useful in Chapter 6.

2.1 Unbounded operators

Throughout this chapter $\mathcal{H}$ will denote a complex, separable Hilbert space. We assume that the inner product on $\mathcal{H}$ is linear in the first argument and conjugate linear in the second. Let $\mathcal{L}(\mathcal{H})$ denote the vector space of linear operators mapping H into itself. We will always assume that the domain $\mathcal{D}(T)$ of an operator $T \in \mathcal{L}(\mathcal{H})$ is a dense subspace of $\mathcal{H}$. For any $T \in \mathcal{L}(\mathcal{H})$ its *operator norm* is given by:

$$\|T\| \coloneqq \sup\{\|Tx\| \colon x \in \mathcal{D}(T), \|x\| = 1\}.$$

As usual, $\mathcal{B}(\mathcal{H}) \coloneqq \{T \in \mathcal{L}(\mathcal{H}) \colon \|T\| < \infty\}$. We say that T is *unbounded* if its operator norm is not finite. Unlike for bounded operators, which are usually defined on the entire space $\mathcal{H}$, the domain of an operator T plays a central role in its definition. Going forward, the term *operator* will refer to a linear operator which is not necessarily bounded.

Graph, closure and core

In the study of general operators, identifying operators with their *graphs* has turned out to be useful. Throughout, we let $T \in \mathcal{L}(\mathcal{H})$ be an arbitrary operator, possibly unbounded.

Definition 2.1. (Graph.) The *graph* of T, denoted $\mathcal{G}(T)$ is defined as the set of pairs,

$$\mathcal{G}(T) \coloneqq \{(x, Tx) \colon x \in \mathcal{D}(T)\}.$$

We view $\mathcal{G}(T)$ as a subspace of the Hilbert space $\mathcal{H}\times\mathcal{H}$ with the component-wise inner product $\langle(x,y),(x',y')\rangle = \langle x,x'\rangle + \langle y,y'\rangle$. ▲

Definition 2.2. (Closedness.) An operator is called *closed* if $\mathcal{G}(T)$ is a closed subset of $\mathcal{H}\times\mathcal{H}$. This is equivalent to the following condition: If $x_n \to x$ and $Tx_n \to y$ where $x_n \in \mathcal{D}(T)$ then $x \in \mathcal{D}(T)$ and $Tx = y$. It is important to note that this condition is strictly weaker than that of continuity (boundedness) since if T is continuous then $x_n \to x$ automatically implies $Tx_n \to Tx$. ▲

One of the fundamental problems with unbounded operators is illustrated by the following simple but important fact which is a consequence of the Closed Graph Theorem [24, p. 84]:

Theorem 2.3. *(Hellinger-Toeplitz.) Let $A \in \mathcal{L}(\mathcal{H})$ be an everywhere defined operator acting on the Hilbert space $\mathcal{H}$ which is symmetric, i.e. $\langle Ax,y\rangle = \langle x,Ay\rangle$ for all $x,y\in\mathcal{H}$. Then A is bounded.*

As a consequence, if we have an unbounded operator which is symmetric, it cannot be defined on all of $\mathcal{H}$. Since many unbounded operators are naturally symmetric where they are defined, Theorem 2.3 suggests why in order to define an unbounded operator we must first decide its domain, $\mathcal{D}(T)$ and then how it acts on elements of $\mathcal{D}(T)$. Most operators of interest will also be closed or *closable*:

Definition 2.4. (Closable.) T is *closable* if it there exists a closed operator T_1 that *extends* T, meaning that $\mathcal{G}(T)\subset\mathcal{G}(T_1)$ or equivalently that $\mathcal{D}(T)\subset\mathcal{D}(T_1)$ and $T_1x = Tx$ for all $x\in\mathcal{D}(T)$. We then write $T\subset T_1$. If T is closable then the minimal closed extension of T is called its *closure* and is denoted by $\overline{T}$. ▲

A natural attempt to obtain a closed operator from T would be to consider the closure of the graph $\mathcal{G}(T)$. However, $\overline{\mathcal{G}(T)}$ is not necessarily the graph of a linear operator anymore. The closable property ensures that this holds:

Proposition 2.5. *If T is closable then $\overline{\mathcal{G}(T)} = \mathcal{G}(\overline{T})$.*

Proof. Consider an arbitrary closed extension S of T. Clearly $\overline{\mathcal{G}(T)}\subset\mathcal{G}(S)$. Define an operator R with domain $\mathcal{D}(R) = \{x\in\mathcal{H} : (x,y)\in\overline{\mathcal{G}(T)} \text{ for some } y\in\mathcal{H}\}$ by $Rx = y$ i.e. the unique vector such that $(x,y)\in\overline{\mathcal{G}(T)}$ (this vector exists and is unique since $\overline{\mathcal{G}(T)}$) is a subset of the graph $\mathcal{G}(S)$). By definition of R, $\mathcal{G}(R) = \overline{\mathcal{G}(T)}$. Since S is an arbitrary closed extension of T, we must have $R = \overline{T}$. ■

It is often convenient to not work with T on all of its domain, but rather on an appropriate dense subset of $\mathcal{D}(T)$ where the operator behaves nicely, such as $C_c^\infty(\mathbb{R})$ for differential operators. This motivates the next definition which will play an important role later:

Definition 2.6. (Core.) A subset $D\subset\mathcal{D}(T)$ is called a *core* for T if for each $x\in\mathcal{D}(T)$ then there is a sequence $\{x_n\}_{n=1}^\infty \subset D$ such that $x_n\to x$ and $Tx_n\to Tx$. If T is closed, then this is equivalent to $\overline{T\restriction D} = T$ where $T\restriction D$ is the restriction of T to D. ▲

Note that $D\subset\mathcal{D}(T)$ being a core for T is equivalent to D being dense in $\mathcal{D}(T)$ endowed with the *graph norm* $\|x\|_T := \|x\| + \|Tx\|$. If T is bounded and hence continuous, then it is clear than any dense subset of $\mathcal{D}(T)$ is a core for T.

Symmetry and self-adjointness

Let $\mathcal{C}(\mathcal{H})$ denote the set of densely defined, closed operators on $\mathcal{H}$. Clearly $\mathcal{B}(\mathcal{H}) \subset \mathcal{C}(\mathcal{H})$, but whereas $\mathcal{B}(\mathcal{H})$ is a C^*-algebra, the set $\mathcal{C}(\mathcal{H})$ is not even a vector space since operators cannot be added in a straightforward manner due to difference of domains. For example, defining the adjoint requires some care to be taken (note that for T^*y to be unique, we need $\mathcal{D}(T)$ to be dense):

Definition 2.7. (Adjoint/Self-adjointness.) Let $T : \mathcal{D}(T) \to \mathcal{H}$ be a densely defined operator. Its adjoint $T^* : \mathcal{D}(T^*) \to \mathcal{H}$ has domain $\mathcal{D}(T^*)$ consisting of all $y \in \mathcal{H}$ such that there exists a vector w satisfying

$$\langle Tx, y\rangle = \langle x, w\rangle$$

for all $x \in \mathcal{D}(T)$. For each such y we define $T^*y = w$. If T is symmetric and $\mathcal{D}(T^*) \subset \mathcal{D}(T)$ then $T = T^*$ and T is called *self-adjoint*. ▲

Note that $T \subset S$ implies $S^* \subset T^*$.

Example 2.8. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and let $\mathcal{H} = L^2(\Omega)$. Fix a measurable function $\psi : \mathbb{R} \to \mathbb{C}$ that is finite μ-almost everywhere. Let M_ψ be the multiplication operator by ψ defined by

$$\begin{aligned}(M_\psi f)(t) &= \psi(t) \cdot f(t), \quad f \in L^2(\Omega),\\ \mathcal{D}(M_\psi) &= \{f \in L^2(\Omega) \colon \psi \cdot f \in L^2(\Omega)\}.\end{aligned}$$

We will show that $M_{\overline{\psi}} = (M_\psi)^*$. To show that $\mathcal{D}(M_\psi)$ is dense, consider arbitrary $f \in L^2(\Omega)$, and define the measurable sets $A_n \coloneqq \{t \in \Omega : |\psi(t)| \le n\}$ with characteristic function χ_n for $n \in \mathbb{N}$. Then clearly $\chi_n f \in \mathcal{D}(M_\psi)$ for each n since

$$\int_\Omega |\psi \chi_n f|^2 \,\mathrm{d}\mu = \int_{A_n} |\psi|^2 |f|^2 \,\mathrm{d}\mu \le n^2 \|f\| < \infty.$$

By assumption, χ_n converges a.e. to the function $x \mapsto 1$ on Ω and so $\chi_n f$ converges to f μ-a.e. and hence $|\chi_n f - f|^2 \to 0$ almost everywhere. By the Dominated Convergence Theorem, $\|\chi_n f - f\| \to 0$ which shows that $\mathcal{D}(M_\psi)$ is dense in $\mathcal{H}$. Next, for any $f, g \in \mathcal{D}(M_\psi)$, clearly

$$\langle M_\psi f, g\rangle = \int_\Omega \psi\, f\, \bar{g} \,\mathrm{d}\mu = \int_\Omega f\, \overline{\overline{\psi} g} \,\mathrm{d}\mu = \langle f, M_{\overline{\psi}} g\rangle$$

so $M_{\overline{\psi}} \subseteq (M_\psi)^*$. For the other inclusion let $g \in \mathcal{D}((M_\psi)^*)$ and set $h = (M_\psi)^* g \in L^2(\Omega)$. We want to show that $g \in \mathcal{D}(M_{\overline{\psi}})$ and that $M_{\overline{\psi}} g = \overline{\psi} g = h$.

For any $f \in \mathcal{D}(M_\psi)$, letting $f_n = \chi_n f$ we have $\langle M_\psi f_n, g\rangle = \langle \psi \chi_n f, g\rangle = \langle \chi_n f, h\rangle$ which means

$$\int_\Omega f \chi_n \left(\psi \bar{g} - \bar{h}\right) = 0.$$

The above holds for all $n \in \mathbb{N}$ and for a finitely large n we have $\chi_n(t) = 1$ for almost all t. Since f is arbitrary and $\mathcal{D}(M_\psi)$ is dense, the function $\chi_n \left(\psi \bar{g} - \bar{h}\right)$ must equal zero a.e. on Ω for all n and so $\psi \bar{g} - \bar{h} = 0$ in $L^2(\Omega)$ i.e. $\overline{\psi} g = h$. △

Independent of T, the operator T^* is always closed. In order to see this, suppose that $y_n \to y$ where $y_n \in \mathcal{D}(T^*)$ and $T^* y_n \to z$. By Definition 2.7, for each $x \in \mathcal{D}(T)$ we have

$$\langle Tx, y_n\rangle = \langle x, T^* y_n\rangle.$$

By continuity of the inner product, taking $n \to \infty$ on both sides gives

$$\langle Tx, y\rangle = \langle x, z\rangle.$$

Since this holds for every $x \in \mathcal{D}(T)$, we have $y \in \mathcal{D}(T^*)$ and $T^*y = z$, showing that $(y, T^*y) \in \mathcal{G}(T^*)$, hence the graph of T^* is closed. Recall the following fact from Hilbert space theory:

Proposition 2.9. *Let U be a unitary map, i.e., $U^*U = UU^* = I$ on $\mathcal{H}$. Then U commutes with orthogonal complements in the sense that for any closed subspace M of $\mathcal{H}$, $U(M^\perp) = U(M)^\perp$.*

Proof. This follows directly from the fact that for $x \in M^\perp$ and $y \in M$ we have

$$\langle Ux, Uy\rangle = \langle U^*Ux, y\rangle = \langle x, y\rangle = 0.$$

■

Proposition 2.10. *An operator T is closable if and only if $\mathcal{D}(T^*)$ is dense. If it is closable then $\overline{T} = T^{**}$.*

Proof. First define the unitary map U acting on $\mathcal{H}\times\mathcal{H}$ given by $(x, y) \mapsto (y, -x)$ and note that for any operator A,

$$\begin{aligned}(y, w) \in U(\mathcal{G}(A))^\perp &\iff \langle (y, w), (-Ax, x)\rangle = 0 \;\forall\; x \in \mathcal{D}(A)\\ &\iff \langle Ax, y\rangle = \langle x, w\rangle \;\forall\; x \in \mathcal{D}(A)\\ &\iff (y, w) \in \mathcal{G}(A^*),\end{aligned}$$

i.e. $\mathcal{G}(A^*) = U(\mathcal{G}(A))^\perp$. Moreover, observe that since U^2 is also unitary and commutes with orthogonal complements,

$$\begin{aligned}\overline{\mathcal{G}(A)} &= \mathcal{G}(A)^{\perp\perp}\\ &= (U^2(\mathcal{G}(A)^\perp))^\perp\\ &= (U(U\mathcal{G}(A))^\perp)^\perp\\ &= (U\mathcal{G}(A^*))^\perp.\end{aligned}$$

Thus if $\mathcal{D}(T^*)$ is dense so that T^{**} makes sense, we can use $A = T^*$ in the two above observations to see that $\mathcal{G}(T^{**}) = (U\mathcal{G}(T^*))^\perp = \overline{\mathcal{G}(T)}$ and since T is closable, this means $T^{**} = \overline{T}$. Conversely, if $\mathcal{D}(T^*)$ is not dense, we can pick non-zero $x \in \mathcal{D}(T^*)^\perp$. Then for any $y \in \mathcal{D}(T^*)$ we see that

$$\langle (x, 0), (y, T^*y)\rangle = 0$$

which means that $(x, 0) \in \mathcal{G}(T^*)^\perp$. But then $(0, -x) \in U(\mathcal{G}(T^*))^\perp = \overline{\mathcal{G}(T)}$ so that this set is not the graph of a linear operator, which shows that T is not closable. ■

This gives a simple characterisation of adjoints and inverses in terms of graphs.

Proposition 2.11. *Suppose that $T \in \mathcal{C}(H)$ has a densely defined inverse $T^{-1} \in \mathcal{L}(\mathcal{H})$. Then $(T^{-1})^* = (T^*)^{-1}$.*

Proof. First note that T^{-1} is closed since

$$\mathcal{G}(T^{-1}) = \{(y, T^{-1}) \colon y \in T(\mathcal{H})\} = \{(Tx, x) \colon x \in \mathcal{D}(T)\} = \tau(\mathcal{G}(T)),$$

so $\mathcal{G}(T^{-1})$ is the image of the closed set $\mathcal{G}(T)$ under the unitary *transposition* map $\tau(x, y) = (y, x)$ on $\mathcal{H} \times \mathcal{H}$. By assumption, T^{-1} has a densely defined adjoint $(T^{-1})^*$. We need to check that T^* has an inverse, or equivalently verify that $\tau(\mathcal{G}(T^*))$ is the graph of a linear operator. Recall the unitary map U from Proposition 2.10 which satisfies $\mathcal{G}(S^*) = U[\mathcal{G}(S)]^\perp$ for any $S \in \mathcal{C}(H)$. Noting that $\tau U \mathcal{G}(T) = U \tau \mathcal{G}(T)$, by Proposition 2.9 we have

$$\tau(\mathcal{G}(T^*)) = \tau(U[\mathcal{G}(T)]^\perp) = U[\tau(\mathcal{G}(T))]^\perp = U[\mathcal{G}(T^{-1})]^\perp = \mathcal{G}((T^{-1})^*).$$

The set $\mathcal{G}((T^{-1})^*)$ is the graph of the adjoint of T^{-1}, which we know is a well defined linear operator. Hence the above shows that T^* is indeed an invertible operator with $(T^{-1})^* = (T^*)^{-1}$. ■

Another simple consequence of Proposition 2.10 is that if T is closable then

$$T^* = \overline{(T^*)} = T^{***} = \overline{T}^*.$$

If the stronger condition $T = T^*$ is satisfied, then T is self-adjoint. In the case $\mathcal{D}(T) = \mathcal{H}$ then symmetry implies $T = T^*$ so we see that this generalizes the bounded case. By definition, a symmetric operator is always closable since T^* is automatically closed and it extends T. Many operators of interest are not strictly speaking self-adjoint, but are almost self-adjoint in the following sense:

Definition 2.12. A symmetric operator T is called *essentially self-adjoint* if $\overline{T}$ is self-adjoint. ▲

If T is essentially self-adjoint then $\overline{T}$ is its *only* self-adjoint extension. For suppose S is self-adjoint with $T \subset S$. By Proposition 2.10, $\overline{T} = T^{**}$ and so $T^{**} \subset S$ since S is closed. Thus $S = S^* \subset (T^{**})^* = \overline{T}^*$. An important consequence of this is that if $D \subset \mathcal{D}(T)$ and the restriction $T \restriction D$ is essentially self-adjoint then D is a core for the self-adjoint operator $\overline{T}$. One of the quintessential examples comes from physics, see [30, p. 108]:

Example 2.13. The Hamiltonian

$$H = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + x^2$$

of the *quantum harmonic oscillator* acting on $L^2(\mathbb{R})$ is essentially self-adjoint when defined on the Schwartz space $\mathcal{S}(\mathbb{R})$ of rapidly decreasing smooth functions. Its eigenfunctions are the *Hermite functions* (see Equation (5.11)), which form an orthonormal basis for $L^2(\mathbb{R})$, with eigenvalues $\{2n + 1\}_{n \in \mathbb{N}}$. △

A differentiation operator

Next we consider one of the paradigmatic cases of an unbounded operator: Differentiation. First we recall a definition from real analysis: Let $[a, b] \subset \mathbb{R}$ be a compact interval. A function $f : [a, b] \to \mathbb{C}$ is called *absolutely continuous*

if there exists a Lebesgue integrable function g on $[a, b]$ and a constant $\gamma \in \mathbb{C}$ such that for all $x \in [a, b]$, we have:

$$f(x) = \int_a^x g(t)\,\mathrm{d}t + \gamma.$$

We denote the set of absolutely continuous functions on $[a, b]$ by $AC[a, b]$. If $f \in AC[a, b]$, then f is differentiable almost everywhere (a.e.) and its derivative which we denote f' satisfies $f'(t) = g(t)$ for almost all t. Since each $f \in AC[a, b]$ is bounded on the compact interval $[a, b]$, we have $f \in L^2[a, b]$. However, its derivative need not be in $L^2[a, b]$, consider $f(x) = x^p$ with $0 < p < 1$ on $[0, 1]$.

Next we give an example which illustrates several of the concepts from the previous section: An unbounded, symmetric operator with an adjoint that extends it to a strictly larger subspace.

Remark 2.14. It may seem strange to impose boundary conditions on a set of measure zero in the space L^2, but as long as we restrict ourselves to (equivalence classes of) functions in $AC[0, 1]$ this does make sense due to the regularity of the functions.

Example 2.15. Consider the Hilbert space $\mathcal{H} = L^2[0, 1]$. Since we want an operator with range inside $\mathcal{H}$ we define our operator T by $Tf(x) = if'(x)$ on the domain

$$\mathcal{D}(T) = \{f \in AC[0, 1] : f' \in L^2[0, 1], f(0) = f(1) = 0\}.$$

First we show that T is unbounded and symmetric. First, note that any polynomial f can be approximated to arbitrary accuracy in the L^2-norm by a function in $\mathcal{D}(T)$. To see this, for any $\varepsilon > 0$, pick $\delta > 0$ to cut off the polynomial at points $a_0 = a + \delta$ and $b_0 = b - \delta$ by instead drawing a straight line from $(0, 0)$ to $(a_0, f(a_0))$ and from $(b_0, f(b_0))$ to $(1, 0)$. The modified function f_0 will be in $AC[0, 1]$, and by choosing δ sufficiently close to zero, it will satisfy $\|f - f_0\| < \varepsilon$. Since the polynomials are dense in $L^2[0, 1]$, so is $\mathcal{D}(T)$. To show that T is unbounded, consider the "tent" functions

$$f_n(x) = \begin{cases} nx, & 0 < x \leq 1/n \\ -nx + 2, & 1/n < x \leq 2/n \\ 0, & 2/n < x \leq 1, \end{cases}$$

shown in Figure 2.1. Straightforward calculation gives $\|f_n\|^2 = \frac{2}{3n}$ and $\|Tf_n\|^2 = \|if_n'\|^2 = \int_0^{2/n} n^2\,\mathrm{d}x = 2n$ so that

$$\frac{\|Tf_n\|}{\|f_n\|} = \sqrt{\frac{3n \cdot 2n}{2}} \geq n,$$

hence

$$\|T\| = \sup_{f \neq 0, f \in \mathcal{D}(T)} \frac{\|Tf\|}{\|f\|} = \infty.$$

To see that T is symmetric, i.e. that $\langle Tf, g\rangle = \langle f, Tg\rangle$ for all $f, g \in \mathcal{D}(T)$, integrate by parts:

$$\begin{aligned}\langle Tf, g\rangle &= \langle if', g\rangle \\ &= i\int_0^1 f'(x)\overline{g(x)}\,\mathrm{d}x \\ &= -i\int_0^1 f(x)\overline{g'(x)}\,\mathrm{d}x \\ &= \langle f, Tg\rangle,\end{aligned}$$

where the boundary terms vanish due to the second condition on the domain. Since T is symmetric on $\mathcal{D}(T)$, its adjoint T^* is an extension of T. Next, we

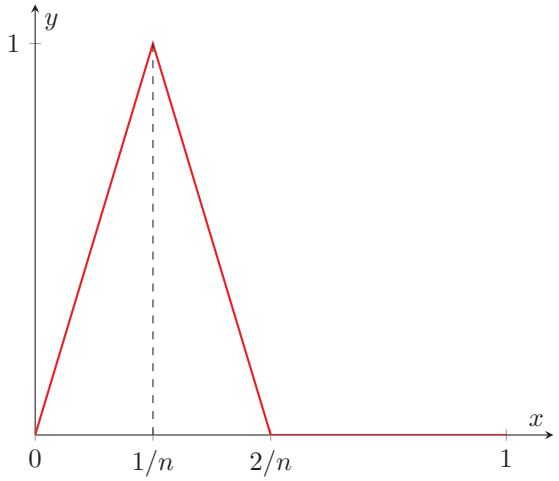


Figure 2.1: The graph of f_n.

will show that T^* is a *proper* extension of T. Specifically, we will show that $T^*f = if'$ on

$$D^* \coloneqq \{f \in AC[0,1] : f' \in L^2[0,1]\}.$$

and that $\mathcal{D}(T^*) = D^*$. For the inclusion $D^* \subset \mathcal{D}(T^*)$, let $g \in \mathcal{D}(T)$ and recall Definition 2.7 which says that we must show that there exists $h \in \mathcal{H}$ such that for any $f \in \mathcal{D}(T)$ we have $\langle Tf, g\rangle = \langle f, h\rangle$, in which case $T^*g \coloneqq h$. But this follows from integration by parts as above since $f(0) = f(1) = 0$ and so,

$$\langle Tf, g\rangle = \int_0^1 if'(x)\overline{g(x)}\,\mathrm{d}x = -\int_0^1 if(x)\overline{g'(x)}\,\mathrm{d}x = \langle f, ig'\rangle.$$

Hence $g \in \mathcal{D}(T^*)$ and $T^*g = ig'$, where $g' \in \mathcal{H}$ by definition of D^*. If we can show the converse inclusion $\mathcal{D}(T^*) \subset D^*$, we are done.

To this end, let $g \in \mathcal{D}(T^*)$ and note that g and T^*g are also integrable since $L^2[0,1] \subset L^1[0,1]$ (by e.g. Hölder's inequality). We want to show that g is absolutely continuous and that $T^*g = ig'$ (which implies $g' \in L^2[0,1]$). In order to do this, consider the absolutely continuous (and hence integrable) function

$$\xi(t) \coloneqq \int_0^t T^*g(s)\,\mathrm{d}s + \mathcal{G}$$

for some constant $\mathcal{G}$ to be chosen later. Then $\xi'(t) = T^*g(t)$ for almost all $t \in [0,1]$. To show that $T^*g = ig'$, the idea is to show that $g(t)$ equals the absolutely continuous function $-i\xi(t)$, which then gives $ig'(t) = \xi'(t) = T^*g(t)$.

By definition of the adjoint, for any $f \in \mathcal{D}(T)$ we have $\langle Tf, g\rangle = \langle f, T^*g\rangle$, or in other words:

$$\int_0^1 if'(t)\overline{g(t)}\,\mathrm{d}t = \int_0^1 f(t)\overline{D^*g(t)}\,\mathrm{d}t = -\int_0^1 f'(t)\overline{\xi(t)}\,\mathrm{d}t,$$

where the second equality follows from integration by parts. This implies that

$$\int_0^1 f'(t)(\overline{g(t) + i\xi(t)})\,\mathrm{d}t = 0. \tag{2.1}$$

This suggests defining a particular $f \in \mathcal{D}(T)$ such that $f' = g + i\xi$. The canonical choice is

$$f(t) = \int_0^t (g(s) + i\xi(s))\,\mathrm{d}s.$$

Then f is absolutely continuous with $f'(t) = g(t) + i\xi(t)$ for almost all $t \in [0,1]$. Clearly $f(0) = 0$. Finally, to ensure that $f(1) = 0$ note that by definition of ξ (recall g and T^*g are integrable):

$$f(1) = \int_0^1 g(t)\,\mathrm{d}t + i\int_0^1\int_0^t T^*g(s)\,\mathrm{d}s\,\mathrm{d}t + i\mathcal{G}.$$

and so we choose

$$\gamma = i\int_0^1 g(t)\,\mathrm{d}t - \int_0^1\int_0^t T^*g(s)\,\mathrm{d}s\,\mathrm{d}t.$$

Then $f \in \mathcal{D}(T)$ and (2.1) gives $\|g + i\xi\| = 0$, i.e., $g = -i\xi$ which finally implies $g' = -i\xi' = -iT^*g \in L^2[0,1]$ as we wanted to show.

△

This rather lengthy example illustrates the inherent subtleties in constructing an unbounded self-adjoint operator. To get the symmetry property one needs to restrict the domain, but it cannot be "too small", since then T^* will extend T properly.

2.2 The spectrum of unbounded operators

The spectrum of an operator in a Hilbert space generalises the idea of eigenvalues for a matrix from finite dimensional linear algebra. As in linear algebra, an eigenvalue of an operator $T \in \mathcal{L}(\mathcal{H})$ is a complex number λ such that $T - \lambda I$ has a non-trivial kernel, where I is the identity operator on $\mathcal{H}$. If $\mathcal{H}$ is finite dimensional, injectivity and surjectivity are equivalent, so the above is equivalent to $T - \lambda I$ not having an inverse in $\mathcal{B}(\mathcal{H})$. In the infinite dimensional case however, even if $T - \lambda I$ is injective and we can formally make sense of $(T - \lambda I)^{-1}$, other problems may arise:

- The domain of $(T - \lambda I)^{-1}$ is not equal to the entire space $\mathcal{H}$, i.e. the image $\mathcal{R}(T - \lambda I)$ of $T - \lambda I$ is not all of $\mathcal{H}$.

- $(T - \lambda I)^{-1}$ may not be bounded.

This motivates the following definition:

Definition 2.16. (Spectrum and resolvent.) Let $T \in \mathcal{C}(\mathcal{H})$. The *resolvent set* $\rho(T)$ of T is the set of all $z \in \mathbb{C}$ such that $T - zI$ is a bijection of $\mathcal{D}(T)$ onto $\mathcal{H}$ with a bounded inverse, which we denote by $R(z, T) \coloneqq (T - zI)^{-1}$ and refer to as the *resolvent operator of T at z*. The *spectrum* of T, denoted by $\mathrm{Sp}(T)$ is the complement of the resolvent set, $\mathrm{Sp}(T) = \mathbb{C} \setminus \rho(T)$. ▲

Given the different ways in which $T - zI$ may fail to have an inverse that is an element of $\mathcal{B}(\mathcal{H})$, the spectrum can be decomposed into several disjoint parts. We restrict the attention to self-adjoint operators and refer to [26] for more details.

Definition 2.17. (Discrete, essential spectrum.) Let T be self-adjoint. Its *discrete spectrum* $\mathrm{Sp}_{\mathrm{disc}}(T)$ is the set of all eigenvalues of T which are isolated points of $\mathrm{Sp}(A)$ and whose corresponding eigenspace is finite dimensional. The *essential spectrum* $\mathrm{Sp}_{\mathrm{ess}}(T)$ of T is the complement of $\mathrm{Sp}_{\mathrm{disc}}(T)$ in $\mathrm{Sp}(T)$. ▲

Remark 2.18. Which of the disjoint parts of the spectrum each $z \in \mathrm{Sp}(T)$ belongs to, corresponds to "how badly" $T - zI$ fails to be invertible. If T is not self-adjoint, there are several different definitions of $\mathrm{Sp}_{\mathrm{ess}}(T)$ but they all coincide with Definition 2.17 when T is self-adjoint. There is also the notion of the *residual spectrum*, the set of all $z \in \mathrm{Sp}(T)$ such that z is not an eigenvalue and $\mathcal{R}(T - zI)$ is not dense in $\mathcal{H}$, but this is empty for self-adjoint operators [24, Theorem VI.8]

Just as in the bounded case, if T is self-adjoint we have $\mathrm{Sp}(T) \subset \mathbb{R}$ [26, Section 3.2]. We recall that for any $T \in \mathcal{B}(\mathcal{H})$, the spectrum of T is a non-empty and compact subset of $\mathbb{C}$ [24, pp. 188–191]. The proof that the spectrum for a general $T \in \mathcal{B}(\mathcal{H})$ is non-empty relies on Liouville's theorem and is a proof by contradiction. A fundamental observation is that the proof does not provide any indication as to how one may actually find any points in the spectrum, as pointed out in [9]. By a standard Neumann series argument, the spectrum of an unbounded operator is always closed in $\mathbb{C}$, but it need no longer be bounded.

Next, we give consider an example which shows that the choice of which dense domain to use is not an annoying technicality, but a fundamental property of an unbounded operator [24].

Example 2.19. Consider the operator $Tf = if'$ from Example 2.15 acting on the subspace $AC[0, 1] \subset L^2[0, 1]$ of absolutely continuous functions, and define the two domains,

$$D_1 \coloneqq \{f \in AC[0, 1] : f' \in L^2[0, 1]\}$$
$$D_2 \coloneqq \{f \in AC[0, 1] : f' \in L^2[0, 1], f(0) = 0\}.$$

First consider $\mathcal{D}(T) = D_1$. Given any $\lambda \in \mathbb{C}$, consider the function $f \in D_1$ defined by $f(x) = e^{-i\lambda x}$, which gives $Tf = \lambda f$. Thus $\mathrm{Sp}(T) = \mathbb{C}$. Next consider the case $\mathcal{D}(T) = D_2$. Given $\lambda \in \mathbb{C}$ define the operator S_λ by

$$(S_\lambda g)(x) = -i \int_0^x e^{-i\lambda(x-s)} g(s)\, \mathrm{d}s.$$

For $g \in L^2[0,1]$ we compute

$$\frac{\mathrm{d}}{\mathrm{d}x}(S_\lambda g)(x) = -i\left[\lambda(S_\lambda g)(x) + g(x)\right]$$
$$(T-\lambda I)(S_\lambda g)(x) = \left(i\frac{\mathrm{d}}{\mathrm{d}x} - \lambda\right)(S_\lambda g)(x) = g(x),$$

and so $(T-\lambda I)(S_\lambda g) = I$ on $L^2[0,1]$. Integration by parts and the boundary condition on D_2 shows that $S_\lambda(T-\lambda I)g = g$ for $g \in D_2$. So for any $\lambda \in \mathbb{C}$, the operator S_λ is an inverse to $T-\lambda I$. Finally, it is easily checked that $\|S_\lambda g\| \leq C\|g\|$ for all $g \in L^2[0,1]$ where C is some constant only depending on λ. Thus $T-\lambda I$ has a bounded inverse S_λ for every $\lambda \in \mathbb{C}$ and so $\mathrm{Sp}(T) = \emptyset$. $\triangle$

Next we have a very simple but important observation that is a first step towards locating the spectrum of a given operator. Given any non-empty closed set $X \in \mathbb{C}$ and $z \in \mathbb{C}$ we define $\mathrm{dist}(z, X) \coloneqq \inf_{x\in X}|z-x|$.

Proposition 2.20. *Let T be a closed operator. Then for each $z \in \rho(T)$ we have*

$$\|R(z,T)\|^{-1} \leq \mathrm{dist}(z, \mathrm{Sp}(T)).$$

Proof. Let $z \in \rho(T)$ and suppose that $w \in \mathbb{C}$ satisfies $|w-z| < \|R(z,T)\|^{-1}$. Then $\|(w-z)(T-zI)^{-1}\| < 1$ and by the standard Neumann series argument,

$$I - (w-z)(T-zI)^{-1}$$

is invertible. But then so is

$$(T-zI)(I-(w-z)(T-zI)^{-1}) = T - zI - (w-z)I = T - wI,$$

meaning that $w \in \rho(T)$. Thus if $w \in \sigma(T)$ we must have $|w-z| \geq \|R(z,T)\|^{-1}$ as claimed. ■

This simple observation shows that the norm of the resolvent of T at a point z is explicitly related to the distance from z to the spectrum of T, and this is indeed the idea behind the algorithms which we will develop later. The above holds with equality if T is self-adjoint.

Proposition 2.21. *Let T be unbounded and self-adjoint and $z \in \rho(T)$. Then*

$$\|R(z,T)\|^{-1} = \mathrm{dist}(z, \mathrm{Sp}(T)).$$

Proof. By Proposition 2.11, $R(z,T)^* = (T-\bar{z}I)^{-1}$, so it is easy to see that $R(z,T)$ is a bounded normal operator. Because T is unbounded, we must have $0 \in \mathrm{Sp}(R(z,T))$, since otherwise $R(z,T)^{-1} = T - zI$ would be bounded. Since T is self-adjoint, considering the continuous function $f\colon \mathrm{Sp}(T) \to \mathbb{C}$ defined by $f(w) = \overline{(w-z)^{-1}}$, the spectral mapping theorem for A [26, p. 105] yields $\mathrm{Sp}(f(T)) = \overline{f(\mathrm{Sp}(T))}$, i.e.,

$$\mathrm{Sp}((T-zI)^{-1}) = \left\{\frac{1}{w-z} : w \in \mathrm{Sp}(T)\right\} \cup \{0\},$$

since 0 is the only point on the boundary of $f(\mathrm{Sp}(T))$. Because $(T-zI)^{-1}$ is bounded and normal, its operator norm equals its spectral radius, and the functional calculus for A gives

$$\|(T-zI)^{-1}\| = \sup_{w\in\mathrm{Sp}((T-zI)^{-1})}|w| = \sup_{w\in\mathrm{Sp}(T)}\frac{1}{|z-w|} = \frac{1}{\mathrm{dist}(z,\mathrm{Sp}(T)}.$$

■

2.3 Tensor products of Hilbert spaces

We close this chapter with a description of the tensor product of two Hilbert spaces, in the familiar case L^2, which will be used in later chapters, following [24]. Let (M, μ) and (N, ν) be σ-finite measure spaces so that $L^2(M, d\mu)$ and $L^2(N, d\nu)$ are separable Hilbert spaces. For arbitrary $\varphi_1 \in L^2(M, d\mu)$ and $\psi_1 \in \times L^2(N, d\nu)$, let $\varphi_1 \otimes \psi_1$ denote the conjugate bilinear form acting on $L^2(M, d\mu) \times L^2(N, d\nu)$ by

$$(\varphi_1 \otimes \psi_1)(\varphi_2, \psi_2) \coloneqq \langle \varphi_2, \varphi_1 \rangle_{L^2(M,d\mu)} \cdot \langle \psi_1, \psi_2 \rangle_{L^2(N,d\nu)}.$$

Denote by $\mathcal{E}$ the set of all finite linear combinations of such conjugate bilinear forms, and define an inner product on $\mathcal{E}$ by letting

$$\langle \varphi \otimes \psi, \varphi' \otimes \psi' \rangle \coloneqq \langle \varphi, \varphi' \rangle \cdot \langle \psi, \psi' \rangle \tag{2.2}$$

The tensor product $L^2(M, d\mu) \otimes L^2(N, d\nu)$ is formally defined by taking the completion of $\mathcal{E}$ with respect to the inner product defined in (2.2). Next, assume we have orthonormal bases $\{\varphi_m\}$ and $\{\psi_n\}$, for $L^2(M, d\mu)$ and $L^2(N, d\nu)$ respectively. Then the collection $\{\varphi_m \otimes \psi_n\}$ is an orthonormal basis for $L^2(M, d\mu) \otimes L^2(N, d\nu)$. Letting $d\mu \otimes d\nu$ denote the product measure on $L^2(M \times N)$, it is well known that $L^2(M \times N, d\mu \otimes d\nu)$ is also a Hilbert space. Using Fubini's theorem, the collection $\{\varphi_m \psi_n\}_{m,n}$, where $(\varphi_m \psi_n)(x, y) = \varphi_m(x)\psi_n(y)$ for $(x, y) \in M \times N$, is easily shown to be an orthonormal basis for $L^2(M \times N, d\mu \otimes d\nu)$. Now consider

$$U : \varphi_m \otimes \psi_n \mapsto \varphi_m \psi_n,$$

which maps an orthonormal basis for $L^2(M, d\mu) \otimes L^2(N, d\nu)$ onto an orthonormal basis for $L^2(M \times N, d\mu \otimes d\nu)$. Thus U extends uniquely to a unitary mapping from

$$L^2(M, d\mu) \otimes L^2(N, d\nu) \text{ onto } L^2(M \times N, d\mu \otimes d\nu),$$

and so the spaces are isomorphic.

CHAPTER 3

Computability

In this chapter we introduce the theory of computability. Unlike *complexity theory* from computer science which is concerned with quantifying the efficiency of algorithms, computability theory only concerns itself with the existence of algorithms that solve a given problem in a finite amount of steps, without regard for the number of computational steps needed as long as it is finite.

Section 3.1 provides some historical context and motivation from classical computability à la Gödel and Turing, and can safely be skimmed or skipped altogether by the reader familiar with computability theory. Section 3.2 lays out the more general theory of computation introduced [16] to study the problem of computing spectra. Finally, Section 3.3 shows how this theory can be used to construct algorithms that provide rigorous estimates for spectra of unbounded operators, which will be central to our applications to Dirac operators in later chapters. We will mostly follow the exposition given in [8], often stating slightly simplified (but sufficient for our purposes) versions of the results, with somewhat more elaborate proofs, clarifying a few details.

3.1 Classical computability

Driven by a belief that unsolvable problems did not exist in mathematics, Hilbert initiated his program to clarify the foundations of mathematics in the 1920s. In 1931 however, Kurt Gödel proved his First Incompleteness Theorem which says that any computable axiomatic system that contains basic arithmetic cannot be both *consistent* (only proves true statements) and *complete* (able to prove all true statements) [20, p. 176]. Only a few years after Gödels's breakthrough, Alan Turing came up with the idea of a formal computing machine (today known as the Turing machine).

Turing computation

In this section, a *total* function is a function that is defined on all of $\mathbb{N}$, and this is what is meant by a *computable* function in the following. A *partial computable* (p.c.) function ψ is a function that may be undefined for some inputs. Its domain $\mathrm{dom}(\psi)$, is the subset of $\mathbb{N}$ where ψ is defined.

Following [27], a Turing machine M consists of an infinite *tape* partitioned into *cells* which are either blank (B) or contain the symbol 1 and a *tape head* which scans one cell at a time. It works in discrete time steps and at each step

in time, M is in one of a finite set of *states* Q. There is a special starting state q_1 where the tape head located at the leftmost cell containing a 1, and a final *halting state* q_0 where the machine stops moving and returns an output number. At each step in time the machine can:

1. Change the state of the machine to a different element in Q.
2. Change the symbol of the cell currently being scanned by the tape head to an element in the "alphabet" $A = \{B, 1\}$.
3. Move the tape head one cell left (L), right (R), or let it stay in place (S).

The input x is represented by $x + 1$ consecutive 1's and all other cells blank. How M operates is determined by a function $\delta : Q \times A \to Q \times A \times \{L, S, R\}$, which is a *partial* function, meaning it may be undefined for some inputs. If $\delta(s, q) = (s', q', X)$, then upon scanning a symbol s in state q, the machine M will change its state to q', replace s by s', and move the tape head according to whether $X = L$, $X = S$ or $X = R$. We assume that there are no elements on the form (q_0, s) in the domain of δ, i.e. the machine terminates when the halting state is reached. If M reaches q_0 after t steps with the integer y written on the tape, then we say that M *halts* and outputs y. We say that M computes the partial function $\psi : \mathrm{dom}(\psi) \to \mathbb{N}$ when for each $x \in \mathbb{N}$ we have that $\psi(x) = y$ if and only if M halts and outputs y on input x.

The map δ is called a *Turing program* and corresponds to a finite set of 5-tuples. Thus each Turing program can be identified with a unique finite sequence of numbers. In [27], the author uses a so-called Gödel-numbering which encodes a sequence of natural numbers $(x_1, \ldots, x_n)$ as the product $p_1^{x_n+1} \cdots p_n^{x_n+1}$ where p_n is the n-th prime. This defines a bijective encoding map that can be computed and inverted in finite time. Let P_e be the Turing program whose number sequence is encoded by the Gödel numbering to e. Let ϕ_e denote the partial computable (p.c.) function which is calculated by P_e. We say that e is the *index* of ϕ_e. We say that $\phi_e(x)$ *converges* if and only if the Turing program P_e halts on input x. Clearly each p.c. function has infinitely many indices, since given some P_e we can construct another program $P_{e'}$ which computes the same function by adding extraneous instructions that never affect the output.

Example 3.1. Consider the function $f(x) = 2x$. The input x is represented by a string of $x + 1$ 1's and the output number equals the total number of 1s on the tape upon halting. One idea for a Turing machine that computes $f(x)$ is to delete the first 1 from the input string and then search along the input string, and for each remaining 1, delete it and write two 1's on a string that we build to the left of the input string (maintaining one blank cell in the middle to separate input and output). △

Definition 3.2. A set $A \subset \mathbb{N}$ is called computable if its indicator function is a (total) computable function, and similarily for relations $R \subset \mathbb{N} \times \mathbb{N}$ which can naturally be identified with a p.c. function. Clearly, A is computable if and only if its complement $\overline{A}$ is computable, since the indicator function of A can be trivially computed given the indicator function of $\overline{A}$ and vice versa. ▲

The Church-Turing thesis asserts that a total function $f : \mathbb{N} \to \mathbb{N}$ is *computable* if and only if there is a Turing machine which computes its value on

any input in a finite amount of time. In this spirit, the term *algorithm* is used synonomusly with computable function. When we say that a procedure can be done *efficiently*, we mean that it can be performed by a computable function.

Turing used this idea of computability to give a negative answer to Hilbert's famous Entscheidungsproblem (Decision Problem) [32]: Given a finite set of axioms, is there an algorithm which takes as input a statement and outputs **Yes** if the statement is true and **No** if it is false? This initiated computability theory which seeks to classify functions, sets, relations, predicates and other mathematical objects in terms of their computability.

Undecidable problems and semi-computability

Turing's idea of computability is meant to capture exactly those computations that (in theory) could be carried out to arbitrary precision by a human being with pencil and paper in a finite amount of time, but with no limit on how much paper or ink is consumed [32]. It is not at all obvious that there are computations that are logically impossible to write down in such a way. For example the number $\sin(e^\pi)$ can be computed to arbitrary precision using power series expansion, and any measurable function is computable by approximations from simple functions. The fundamental reason why computability fails is that the set of computable functions is countably infinite. The ubiquitous concept of *diagonalisation* shows us why this is the case.

Example 3.3. The set $K \coloneqq \{x \in \mathbb{N} \colon \phi_x(x) \text{ converges}\}$ is not computable. Suppose K has a computable characteristic function. Then the function

$$f(x) = \begin{cases} \phi_x(x) + 1 & \text{if } x \in K \\ 0 & \text{if } x \notin K, \end{cases}$$

is computable. Pick an index e for f, i.e. assume $f = \phi_e$. Since f is total, $\phi_e(e) = f(e)$ must converge, i.e. $e \in K$. But then $f(e) = \phi_e(e) + 1 \neq \phi_e(e)$, so we have a contradiction. △

So K is uncomputable but this is merely the beginning of the story. It does have a property that makes it highly interesting in computability theory, namely it is *computably enumerable*.

Definition 3.4. A set $A \subset \mathbb{N}$ is *computably enumerable (c.e.)* if it is the domain of some partial computable function. Some texts use the term semi-computable and older literature often uses the term *recursively enumerable*. ▲

This definition is perhaps not the most intuitive, but it turns out to be equivalent to the more obviously useful property that there is a computable function that lists all the elements of the set, i.e. each member of A will be guaranteed to show up in the list in a finite amount of time. Namely, a set A is c.e. if and only if there is a total computable function f such that A equals the range $\text{rng}(f)$ of f (see [27, p. 26] or [20, p. 238] for a proof). Of course, all computable sets are computably enumerable. Another equivalent characterisation of semi-computable sets comes from mathematical logic.

Definition 3.5. A set $A \subset \mathbb{N}$ is on Σ_1-form (abbrev. "A is Σ_1") if A is on the form $\{x : (\exists y)R(x, y)\}$ for some computable relation $R \subset \mathbb{N} \times \mathbb{N}$. The subscript 1 indicates that there is a single existential quantifier. ▲

By Definition 3.2, a computable relation can be identified with a total computable function, so if R is a computable relation then the set $\{x : (\exists y)R(x,y)\}$ is precisely the domain of the corresponding function and vice versa. In other words, A is Σ_1 if and only if it is the domain of a computable function, i.e. Definition 3.4 and Definition 3.5 coincide.

It is easy to see that the set K is computably enumerable: For example it is the domain of the function

$$\psi(x) = \begin{cases} 1 & \text{if } \phi_x(x) \text{ converges} \\ \text{undefined} & \text{otherwise,} \end{cases}$$

since then $\psi(x)$ evaluates to a number if and only if $x \in K$. The existence of noncomputable c.e. sets such as K turns out to be extremely useful in applications to other areas of mathematics. For example, the set K can be used to give relatively simple proofs of the undecidability of Hilbert's Entscheidungsproblem and Gödel's First Incompleteness Theorem [20, pp. 244–253].

The hierarchy of computation

The noncomputable sets which are c.e. do not intuitively seem *that* intractable, since all of their elements can be efficiently listed. One may think of a $A \subset \mathbb{N}$ as being c.e. if there is a Turing machine that given a number $a \in A$ will halt after a finite amount of time and confirm that a is a member of A, but if $a \notin A$, the machine will run forever without resolving the membership question. This raises the question of whether there are any (necessarily noncomputable) sets which are *not* semi-computable. These sets would intuitively be strictly "more" uncomputable than the semi-computable sets, since their elements cannot even be listed. It turns out that such sets do indeed exist.

Example 3.6. The complement $\overline{K}$ of the set K from Example 3.3 is not semi-computable. For suppose that $\overline{K}$ is c.e. Then we can write $K = \text{rng}(f_0)$ and $\overline{K} = \text{rng}(f_1)$ for two total computable functions f_0 and f_1. Now define $f(x)$ as the smallest number i such that $f_0(i) = x$ or $f_1(i) = x$. Since the ranges of f_0 and f_1 are disjoint and have union $\mathbb{N}$, given any x, there will be a finite number i such that exactly one of $f_0(i)$ and $f_1(i)$ equals x. Thus f is a well defined and total computable function. Finally, we can define the computable function

$$\chi(x) = \begin{cases} 1 & \text{if } f_0(f(x)) = x \\ 0 & \text{otherwise.} \end{cases} \tag{3.1}$$

Then $\chi(x) = x = 1$ if $x \in K$ and $\chi(x) = 0$ if $x \in \overline{K}$. In other words, χ is the indicator function of K. Since K is not computable, we have a contradiction, hence $\overline{K}$ cannot be the range of a computable function, i.e. it is not semi-computable. △

In the example above, K could be replaced by any semi-computable but noncomputable set. The argument above shows that if A and $\overline{A}$ are semi-computable then A is computable via the characteristic function in (3.1). This is just a formal way of saying that if we have a Turing machine that lists all the elements of A and a Turing machine that lists the elements of $\overline{A}$, then for any

$x \in \mathbb{N}$ we can decide its membership in a finite amount of time. Conversely, if A is computable then clearly both A and $\overline{A}$ are semi-computable (since they are both computable).

Definition 3.7. A set A is Π_1 if $A = \{x : (\forall y)R(x, y)\}$ for some computable relation $R \subset \mathbb{N} \times \mathbb{N}$. ▲

Now x satisfies $(\forall y)R(x, y)$ if and only if it does *not* satisify $(\exists y)\neg R(x, y)$, where $\neg R$ is the logical negation of R which is computable iff R is. Thus if $A = \{x : (\forall x)R(x, y)\}$ then $\overline{A} = \{x : (\exists y)\neg R(x, y)\}$. Hence A is Π_1 if and only if $\overline{A}$ is Σ_1.

Example 3.8. The set $\overline{K}$ is Π_1. △

One may think of the Σ_1-sets as sets which can be approximated "from below" and the Π_1-sets as sets which can be approximated "from above". To justify this terminology, consider any Σ_1-set A, and a corresponding Turing machine M that lists all of its elements. We then build a set A' that approximates A by starting with the empty set and adding numbers to A' as they occur in the list given by M. Then A' will only contain elements of A, but if A is infinite, this procedure will never give us the full set A. In this sense, we have an approximation from below, since $A' \subset A$. On the other hand, if A is Π_1 we can begin with $A' = \mathbb{N}$ and since $\overline{A}$ is Σ_1 by definition, we can go through a list of the elements of $\overline{A}$ and remove each one from the set A' as they occur. We will never remove an element of A from A', but the only thing we know for certain is that $A' \supset A$, and so we have an approximation from above.

The distinction between computable, c.e. sets and non-c.e. sets is just the beginning of classical computability theory. The Arithmetical Hierarchy further builds on these concepts using *oracle machines* and the *Turing jump* to get an infinite classification theory for the (un)computability of objects defined over the natural numbers [27, Chp. 4].

With some historical background on the theory of computation established, we turn to the problem of computing the spectrum of a linear operator on an infinite dimensional vector space.

3.2 The Solvability Complexity Index

As laid out in Section 3.1, the classical theory of computation is concerned with determining the difficulty of computing functions defined over the natural numbers. Clearly such a theory can be naturally extended to include any problem regarding computation over a set that can be bijectively mapped onto $\mathbb{N}$, such as the rational numbers. It is not obvious, however, how this theory can extended to the real numbers or more general topological spaces, in order to deal with issues like continuity and convergence that are so central to modern mathematics. One proposal comes from "computable analysis". However, this approach has strong assumptions that do not reflect the potential of computations that can actually be achieved [1, 5]. This motivates the introduction of the more general Solvability Complexity Index (SCI) hierarchy.

Computational problems and general algorithms

We start with an informal description and an example as motivation. The key ingredients in the setup are as follows:

1. A set Ω, called the *domain*
2. A set Λ, containing functions from Ω to $\mathbb{C}$, called the *evaluation set*
3. A metric space $(\mathcal{M}, d)$,
4. A map $\Xi : \Omega \to \mathcal{M}$ called the *problem function.*

The set Ω contains the specific objects for which we wish to compute some feature that sits in the metric space $\mathcal{M}$ and is determined by the map Ξ. The functions in Λ contains the information that we are allowed to read from the objects in Ω.

Example 3.9. For example, Ω can be $\mathcal{B}(\mathcal{H})$ for some Hilbert space $\mathcal{H}$ and $\Xi(A) = \mathrm{Sp}(A)$ for $A \in \Omega$. In this case, a natural choice of (M, d) is the metric space consisting of compact subsets of $\mathbb{C}$ with the Hausdorff metric $d = d_H$,

$$d_H(X, Y) \coloneqq \max\left\{ \sup_{x \in X} \mathrm{dist}(x, Y), \sup_{y \in Y} \mathrm{dist}(y, X) \right\}, \tag{3.2}$$

where $\mathrm{dist}(x, Y) = \inf_{y \in Y} |x - y|$ and $\mathrm{dist}(y, X) = \inf_{x \in X} |x - y|$ for $X, Y \in M$ and $(x, y) \in X \times Y$. The evaluation functions can for example be the matrix elements $f_{ij}(A) = \langle Ae_j, e_i \rangle$ where $\{e_n\}_{n=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$. △

Definition 3.10. (Computational Problem.) A computational problem is a collection $\{\Omega, \Xi, \mathcal{M}, \Lambda\}$ as above, such that for $A_1, A_2 \in \Omega$ then $A_1 = A_2$ if and only if $f(A_1) = f(A_2)$ for every $f \in \Lambda$. When $\mathcal{M}$ and Λ are implied by context, we write $\{\Xi, \Omega\}$ for short. ▲

The "if-part" in the requirement of Definition 3.10 is natural since the collection of functions in Λ represents the total sum of knowledge that we have about objects in the domain Ω, and this should be sufficient to tell the objects apart. Given a computational problem, we want to find functions (also called algorithms) that can be used to approximate Ξ.

Definition 3.11. (General Algorithm.) Given a computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$, a *general algorithm* is a map $\Gamma \colon \Omega \to \mathcal{M}$ such that for $A \in \Omega$:

(i) There is a finite subset of evaluations $\Lambda_\Gamma(A) \subset \Lambda$ such that the action of Γ on A is determined only from $\{f(A)\}_{f \in \Lambda_\Gamma(A)}$.

(ii) For every $B \in \Omega$ that satisfies $f(B) = f(A)$ for all $f \in \Lambda_\Gamma(A)$, then $\Lambda_\Gamma(B) = \Lambda_\Gamma(A)$ and hence $\Gamma(B) = \Gamma(A)$ by (i).

▲

Condition (i) says that on any input, only a finite number of evaluations is needed to determine the output of a general algorithm. There is no restriction on the type of operations allowed in a general algorithm. Condition (ii) ensures that the output of Γ on A is not changed if A is replaced with an object if the change does not affect what is read from the evaluation functions in $\Lambda_\Gamma(A)$.

Next we define a *tower of algorithms* in order to define a hierarchy of computations performed by general algorithms, generalising the Arithmetical Hierarchy described in Section 3.1.

Definition 3.12. (Tower of algorithms) Given a computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$, a *tower of algorithms of height k for* $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ is a (finite) collection of sequences of functions

$$\Gamma_{n_k} \colon \Omega \to \mathcal{M},\ \Gamma_{n_k,n_{k-1}} \colon \Omega \to \mathcal{M},\ \ldots,\ \Gamma_{n_k,\ldots,n_1} \colon \Omega \to \mathcal{M},$$

where $n_1, \ldots, n_k \in \mathbb{N}$ and the functions $\Gamma_{n_k,\ldots,n_1}$ at the bottom level are general algorithms as in Definition 3.11. Furthermore, for each $A \in \Omega$,

$$\begin{aligned}
\Gamma_{n_k,\ldots,n_2}(A) &= \lim_{n_1 \to \infty} \Gamma_{n_k,\ldots,n_1}(A) \\
&\vdots \\
\Gamma_{n_k}(A) &= \lim_{n_{k-1} \to \infty} \Gamma_{n_k,n_{k-1}}(A), \\
\Xi(A) &= \lim_{n_k \to \infty} \Gamma_{n_k}(A),
\end{aligned}$$

where the limits mean convergence in $(\mathcal{M}, d)$. We will use the shorthand notation $\{\Gamma_{n_k,\ldots,n_1}\}$ for a tower of height k. ▲

Definition 3.13. (Arithmetic tower) Given a computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ where Λ is countable, we define an *arithmetic tower of height k for* $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ as a tower of algorithms of height k for $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ where the functions $\Gamma_{n_k,\ldots,n_1} \colon \Omega \to \mathcal{M}$ satisfy the following: For each $A \in \Omega$, the mapping $(n_k \ldots, n_1) \mapsto \Gamma_{n_k,\ldots,n_1}(A)$ is *recursive* and $\Gamma_{n_k,\ldots,n_1}(A)$ is a finite string of complex numbers that can be identified with an element of $\mathcal{M}$. ▲

Remark 3.14. Suppose each $f \in \Lambda$ takes values in $\mathbb{Q}$ and Λ is countable. Recall that $\Gamma_{n_k,\ldots,n_1}(A)$ is determined by the finite set of values $\{f(A)\}_{f \in \Lambda_\Gamma(A)} \subset \mathbb{Q}$. By *recursive*, we mean that there is a Turing machine with oracle $\{f(A)\}_{f \in \Lambda}$ that on input $(n_k \ldots, n_1)$ in finite time halts and outputs $\Gamma_{n_k,\ldots,n_1}(A)$.

Having introduced towers of algorithms, we can define the overarching concept of the computational theory: The Solvability Complexity Index (SCI), which was first introduced in [16] for the computational spectral problems.

Definition 3.15. (Solvability Complexity Index) Given a computational problem $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$, it has *Solvability Complexity Index* $\mathrm{SCI}(\Omega, \Xi, \mathcal{M}, \Lambda) = k$ if $k \in \mathbb{N}$ is the smallest possible height of a tower of algorithms for $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$. If there is a general algorithm Γ such that $\Xi = \Gamma$, then define $\mathrm{SCI}(\Omega, \Xi, \mathcal{M}, \Lambda) = 0$ If no tower of any height exists, then we define $\mathrm{SCI}(\Omega, \Xi, \mathcal{M}, \Lambda) = \infty$. ▲

Going forward, we will be interested in algorithms based on arithmetic operations, so we will assume that (towers of) algorithms are arithmetic from now on. In analogy with the Arithmetical hierarchy, the SCI lets us define the SCI-hierarchy:

Definition 3.16. (SCI-hierarchy.) Consider a collection $\mathcal{C}$ of computational problems and let $\mathcal{T}$ be the collection of all towers of algorithms for problems in $\mathcal{C}$. Then we define,

$$\begin{aligned}
\Delta_0 &\coloneqq \{\{\Xi, \Omega\} \in \mathcal{C} \,|\, \mathrm{SCI}(\Xi, \Omega) = 0\} \\
\Delta_{m+1} &\coloneqq \{\{\Xi, \Omega\} \in \mathcal{C} \,|\, \mathrm{SCI}(\Xi, \Omega) \leq m\},\ m \in \mathbb{N}
\end{aligned}$$

and

$$\Delta_1 \coloneqq \{\{\Xi, \Omega\} \,|\, \exists \{\Gamma_n\} \in \mathcal{T} \text{ s.t. } d(\Gamma_n(A), \Xi(A)) \leq 2^{-n} \; \forall A \in \Omega\}.$$

▲

The class Δ_0 are the computational problems which can be solved exactly by a general algorithm, while Δ_1 are the problems for which one may construct a sequence of general algorithms that on each input A converges in $(\mathcal{M}, d)$ to $\Xi(A)$ with full control over the error, coinciding with Turing's idea of computability. Most problems of interest in spectral theory are in Δ_2 or higher. For example, even for infinite diagonal matrices, the problem of computing the spectrum is not in Δ_1 since one can clearly not have error control as in Definition 3.16.

In order to gain a richer theory of computation that contains the intuitive notions of convergence from above and below, more structure is needed on the metric space $\mathcal{M}$. Specifically, the metric d should behave consistently with respect to the partial order of set inclusion on $\mathcal{M}$. This type of structure is crucial if one is to establish a rigorous control of the error on the output of a general algorithm. In Section 3.1 we described the class Σ_1 (resp. Π_1) as the subsets of $\mathbb{N}$ that may be approximated from below (resp. above). If $\mathcal{M}$ is a totally ordered set, then there is an obvious notion of convergence from below and above:

Definition 3.17. Given Definition 3.16, suppose $\mathcal{M}$ is totally ordered. Define

$$\begin{aligned}
&\Sigma_0 = \Pi_0 = \Delta_0, \\
&\Sigma_1 = \{\{\Xi, \Omega\} \in \Delta_2 \,|\, \exists \{\Gamma_n\} \in \mathcal{T} \text{ s.t. } \Gamma_n(A) \nearrow \Xi(A) \forall A \in \Omega\}, \\
&\Pi_1 = \{\{\Xi, \Omega\} \in \Delta_2 \,|\, \exists \{\Gamma_n\} \in \mathcal{T} \text{ s.t. } \Gamma_n(A) \searrow \Xi(A) \forall A \in \Omega\},
\end{aligned}$$

and for $m \in \mathbb{N}$ let

$$\begin{aligned}
\Sigma_{m+1} &= \{\{\Xi, \Omega\} \in \Delta_{m+2} \,|\, \exists \{\Gamma_{n_{m+1},\ldots,n_1}\} \in \mathcal{T} \,\text{s.t.}\, \Gamma_{n_{m+1}}(A) \nearrow \Xi(A) \; \forall A \in \Omega\}, \\
\Pi_{m+1} &= \{\{\Xi, \Omega\} \in \Delta_{m+2} \,|\, \exists \{\Gamma_{n_{m+1},\ldots,n_1}\} \in \mathcal{T} \,\text{s.t.}\, \Gamma_{n_{m+1}}(A) \searrow \Xi(A) \; \forall A \in \Omega\},
\end{aligned}$$

▲

Considering $\mathcal{M} = \{0, 1\}$ in Definition 3.17 we get the SCI-hierarchy for decision problems, and $\mathcal{M} = \mathbb{N}$ includes the classical Arithmetical Hierarchy as a special case.

The SCI hierarchy for spectral computations

In the case of bounded operators, $\Omega = \mathcal{B}(\mathcal{H})$, the metric space $\mathcal{M}$ is the set of non-empty compact subsets of $\mathbb{C}$ with the Hausdorff metric d_H as in Example 3.9. For closed, unbounded operators, $\mathcal{M}$ is just the set of all closed subsets of $\mathbb{C}$. Here the Hausdorff metric is insufficient. For example, consider the sequence of lines $\ell_n = \{(x, x/n) \colon x \in \mathbb{R}\}$ in $\mathbb{R}^2$. If we try to use the definition in (3.2), clearly $d_H(\ell_n, S) = \infty$ if S is the horizontal axis. Hence the lines do not converge to the horizontal axis as they should in any reasonable metric for unbounded sets.

Instead, we will use Attouch-Wets metric [3] (induced by $\mathbb{C}$ with the Euclidean metric) which may be defined by,

$$d_{\mathrm{AW}}(A,B) \coloneqq \sum_{m=1}^{\infty} 2^{-m} \min\left\{1,\ \sup_{|x|<m} |\mathrm{dist}(x,A) - \mathrm{dist}(x,B)|\right\}, \tag{3.3}$$

for non-empty $A, B \subseteq \mathbb{C}$. This may seem arbitrary, but we will shortly give a characterisation in terms of convergence may make things more clear. But first, we give the formulation of the SCI Hierarchy that is appropriate for spectral computations.

Definition 3.18. (SCI Hierarchy (Hausdorff/AW-metric)) Given Definition 3.16, suppose that $\mathcal{M}$ is either the collection of non-empty compact subsets of $\mathbb{C}$ or the collection of non-empty closed subsets of $\mathbb{C}$ with the Hausdorff or Attouch-Wets metric respectively, denoted by d in either case. Define

$$\begin{aligned}
\Sigma_0 &= \Pi_0 = \Delta_0\\
\Sigma_1 &= \{\{\Xi,\Omega\} \in \Delta_2 \,|\, \exists\{\Gamma_n\} \in \mathcal{T}, \{X_n(A)\} \subset \mathcal{M}\,\text{s.t.}\,\Gamma_n(A) \subset X_n(A),\\
&\qquad\qquad \lim_{n\to\infty} \Gamma_n(A) = \Xi(A),\ d(X_n(A),\Xi(A)) \leq 2^{-n}\,\forall A \in \Omega\},\\
\Pi_1 &= \{\{\Xi,\Omega\} \in \Delta_2 \,|\, \exists\{\Gamma_n\} \in \mathcal{T}, \{X_n(A)\} \subset \mathcal{M}\,\text{s.t.}\,\Xi(A) \subset X_n(A),\\
&\qquad\qquad \lim_{n\to\infty} \Gamma_n(A) = \Xi(A),\ d(X_n(A),\Gamma_n(A)) \leq 2^{-n}\,\forall A \in \Omega\}.
\end{aligned}$$

The above inclusions are as sets in $\mathbb{C}$ and $\{X_n(A)\}$ is a sequence in which $X_n(A) \subset \mathbb{C}$ depends on A. As in Definition 3.17, we can inductively define an infinite hierarchy, but we will not need that and defer to [5]. We will often abuse terminology slightly and use the term "Σ_1-algorithm". ▲

Intuitively, in the Hausdorff metric, the class Σ_1 corresponds to convergence from below, since in particular each point in $\Gamma_n(A)$ is at most 2^{-n} away from $\Xi(A)$. The converse need not hold and $\Xi(A)$ may contain points that are far away from $\Gamma_n(A)$. Similarly the class Π_1 captures convergence from above, since in particular no point in $\Xi(A)$ is more than 2^{-n} away from $\Gamma_n(A)$, but $\Gamma_n(A)$ may contain points far away from $\Xi(A)$. To build a Σ_1-algorithm, by taking subsequences, it is clearly sufficient to construct $\Gamma_n(A)$ such that $\Gamma_n(A) \subset \Xi(A) + B_{E_n(A)}$ for computable functions $E_n(A)$ converging to zero.

There is a useful characterisation of convergence in the Attouch-Wets metric, which shows how the Hausdorff metric is naturally extended to unbounded closed sets. For closed non-empty sets $C_n, C \subseteq \mathbb{C}$ we have $d_{\mathrm{AW}}(C_n, C) \to 0$ if and only if $d_K(C_n, C) \to 0$ for any compact set $K \subset \mathbb{C}$ where

$$d_K(C_1, C_2) = d_H(C_1 \cap K, C_2 \cap K),$$

where d_H is the Hausdorff metric, with the convention that the supremum over the empty set is zero. Equivalently, $d_{\mathrm{AW}}(C_n, C) \to 0$ if and only if for any $\delta > 0$ and compact K there is a finite $N \in \mathbb{N}$ such that if $n > N$ then $C_n \cap K \subset C + B_\delta(0)$ and $C \cap K \subset C_n + B_\delta(0)$. Moreover, it is sufficient to consider compact sets on the form $B_m(0)$ for $m \in \mathbb{N}$ large. We denote by $(\mathrm{Cl}(\mathbb{C}), d_{\mathrm{AW}})$ the metric space of non-empty closed subsets of $\mathbb{C}$ with the Attouch-Wets metric.

This provides a useful criterion for determining Σ_1-convergence as in Definition 3.18 for the Attouch-Wets topology [8, Lemma 6.2].

Lemma 3.19. *Suppose we have a domain Ω, a problem function $\Xi\colon \Omega \to (\mathrm{Cl}(\mathbb{C}), d_{\mathrm{AW}})$ and a sequence of arithmetic algorithms Γ_n each with a finite non-empty output set such that $\Gamma_n(A) \to \Xi(A)$ in the Attouch-Wets metric for all $A \in \Omega$. Suppose also that there are functions $E_n\colon \mathbb{C} \to [0,\infty)$ so that $E_n(z)$ is computable for $z \in \Gamma_n(A)$ and which satisfy*

$$\operatorname{dist}(z, \Xi(A)) \leq E_n(z), \quad \forall z \in \Gamma_n(A), \tag{3.4}$$

and for all $m \in \mathbb{N}$,

$$\lim_{n\to\infty} \sup_{z\in\Gamma_n(A)\cap B_m(0)} E_n(z) = 0. \tag{3.5}$$

Then, given $A \in \Omega$ we can compute in finitely many arithmetic operations a sequence of non-negative numbers $b_n \to 0$ such that $\Gamma_n(A) \subset X_n(A) \in \mathcal{M}$ for some set $X_n(A)$ with

$$d_{\mathrm{AW}}(X_n(A), \Xi(A)) \leq b_n \quad \forall n \in \mathbb{N}. \tag{3.6}$$

Thus in the sense of Definition 3.18, by computing a subsequence of b_n if necessary, Γ_n can be converted into a Σ_1-algorithm for computing Ξ.

Proof. Fix $A \in \Omega$. For $m, n \in \mathbb{N}$ define (writing B_r for $B_r(0), r > 0$),

$$a_n^m := \sup_{z\in\Gamma_n(A)\cap B_m} E_n(z).$$

These numbers are computable since $\Gamma_n(A)$ is finite. Then given $m \in \mathbb{N}$, if $z \in \Gamma_n(A) \cup B_m$, (3.4) gives $\operatorname{dist}(z, \Xi(A)) \leq a_n^m$, i.e.,

$$\Gamma_n(A) \cap B_m \subset \Xi(A) + B_{a_n^m}. \tag{3.7}$$

Clearly $\lim_{n\to\infty} a_n^m = 0$ for all $m \in \mathbb{N}$. Next, define the sets

$$X_n^m := \left((\Xi(A) + B_{a_n^m}) \cap B_m\right) \cup \left(\Gamma_n(A) \cup \{z : |z| \geq m\}\right).$$

By (3.7), clearly $\Gamma_n(A) \subset X_n^m$ for all $m \in \mathbb{N}$. If we consider the non-empty finite set $\Gamma_1(A)$, the inclusion (3.7) easily lets us compute a lower bound $m_1 \in \mathbb{N}$ such that $\Xi(A) \cap B_{m_1} \neq \emptyset$. Now suppose $m \geq 4m_1$ and note that $\Xi(A) \cap B_{m_1} \subset X_n^m$ for all n. Consider z with $|z| < \lfloor m/4 \rfloor \leq m/4$. Then in particular there exists $w \in \Xi(A) \cap B_{m_1}$, so $w \in X_n^m$ for all n. By definition of z we then have

$$|w - z| \leq |w| + |z| \leq m_1 + \lfloor m/4 \rfloor \leq m/2,$$

while for any $|y| > m$ we have

$$|y - z| > m - \lfloor m/4 \rfloor \geq m - m/2 = m/2.$$

It follows that the *closest* points in both $\Xi(A)$ and X_n^m to z must lie in B_m. By definition of X_n^m, this implies that for all n we have the two inequalities

$$\operatorname{dist}(z, X_n^m) \leq \operatorname{dist}(z, \Xi(A)) \leq \operatorname{dist}(z, X_n^m) + a_n^m,$$

for arbitrary $m \geq 4m_1$ and $|z| < \lfloor m/4 \rfloor$. Recalling the definition (3.3), for $m \geq 4m_1$ and all $n \in \mathbb{N}$, this gives

$$\begin{aligned} d_{\mathrm{AW}}(X_n^m, \Xi(A)) &\leq \sum_{k=1}^{\lfloor m/4 \rfloor} 2^{-k} a_n^m + \sum_{k=\lfloor m/4 \rfloor + 1}^{\infty} 2^{-k} \\ &\leq a_n^m + 2^{-\lfloor m/4 \rfloor}. \end{aligned} \tag{3.8}$$

To complete the proof we will define a sequence $m(n)$ such that setting $X_n \coloneqq X_n^{m(n)}$ and $b_n \coloneqq a_n^{m(n)} + 2^{-\lfloor m(n)/4 \rfloor}$ yields (3.6). It is clearly sufficient to ensure that $b_n \to 0$ as $n \to \infty$. We construct the sequence $m(n)$ as follows:

1. For $n \leq 4m_1$, set $m(n) = 4m_1$.
2. For $n > 4m_1$ iterate through $k = 4m_1, \ldots, n$:
   (i) If $a_n^k > 2^{-k}$ for each k, set $m(n) = 4m_1$.
   (ii) Otherwise, set $m(n)$ to be the largest k such that $a_n^k \leq 2^{-k}$.

This clearly defines a computable sequence $\{m(n)\}_{n\in\mathbb{N}}$ and by construction $d_{\mathrm{AW}}(X_n, \Xi(A)) \leq b_n$. Given any $k \geq 4m_1$ there can only be a finite number of indices $n \in \mathbb{N}$ such that $a_n^k > 2^{-k}$, for if not we could take a subsequence $a_{n_j}^k$ that does not converge to zero as $n_j \to \infty$, contradicting the assumption (3.5). By our definition of $m(n)$ for $n > 4m_1$, this implies that for all sufficiently large n we have $a_n^{m(n)} \leq 2^{-m(n)}$, and that $\lim_{n\to\infty} m(n) = \infty$ (since no finite number k can be set equal to $m(n)$ for infinitely many n). We conclude that $b_n \to 0$. ■

Remark 3.20. In the algorithm we shall construct, $\Gamma_n(A)$ will always be a finite set and can be assumed to be non-empty, so the hypothesis in Proposition 3.29 is not a restriction.

A classification counterexample

Before constructing the algorithm we will use to produce Σ_1-results, we give an example showing that such a classification is not to be expected even for quite simple operators unless some extra information is provided [5]. In the following example, we do not need to assume the algorithms are arithmetical, they can be of completely general nature as in Definition 3.11.

Example 3.21. In Definition 3.16, let $\mathcal{C}$ be the collection of computational problems $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ on the following form: Ω is the set of linear operators acting on the Hilbert space $\mathcal{H} = \ell^2(\mathbb{N})$ and the evaluation set Λ consists of the matrix evaluation functions $f_{i,j}(A) = \langle Ae_j, e_i\rangle$ with respect to the canonical basis for $\mathcal{H}$. The problem function Ξ is given by $\Xi(A) = \mathrm{Sp}(A)$ for each $A \in \Omega$ and $\mathcal{M}$ is the space of non-empty compact subsets of $\mathbb{C}$ with the Hausdorff metric. Let Ω_C denote the set of compact operators on $\mathcal{H}$. Then in the sense of Definition 3.18 we have $\{\Xi, \Omega_C\} \notin \Sigma_1 \cup \Pi_1$.

Suppose for contradiction that $\{\Xi, \Omega_C\} \in \Sigma_1$, i.e., there exists a sequence of general algorithms $\{\Gamma_n\}$ such that for every $A \in \Omega_C$ we have $\Gamma_n(A) \to \mathrm{Sp}(A)$ with $\Gamma_n(A) \subset \mathrm{Sp}(A) + B_{2^{-n}}(0)$ and the information $\Lambda_{\Gamma_n}(A)$ used to determine $\Gamma_n(A)$ is finite. Define $N(A, n) \coloneqq \max\{i, j \colon f_{i,j} \in \Lambda_{\Gamma_n}(A)\}$. For $k \in \mathbb{N}$ define the matrix A_k by

$$A_k \coloneqq \begin{pmatrix} 1 & & & & 1 \\ & 0 & & & \\ & & \ddots & & \\ & & & 0 & \\ 1 & & & & 1 \end{pmatrix} \in \mathbb{C}^{k\times k}.$$

Denote by A the infinite matrix with A_k in the upper left corner and zero everywhere else. Clearly $\mathrm{Sp}(A) = \{0, 2\}$ and $A \in \Omega_C$. Let $C = \mathrm{diag}\{1, 0, 0, \ldots\}$ so that $\mathrm{Sp}(C) = \{0, 1\}$. We will now choose k to obtain a contradiction. By assumption, there exists n such that $\Gamma_n(C) \cap B_{1/4}(1) \neq \emptyset$. Now set $k > N(C, n)$ in the definition of A. Since the evaluation functions simply provide the matrix elements, in order to have $f_{m,n}(A) \neq f_{m,n}(C)$ we must have $m \geq k$ or $n \geq k$. But by our choice of $k > N(C, n)$ we have for each $f_{i,j} \in \Lambda_{\Gamma_n(C)}$ that $i, j < k$. Hence for every $f_{i,j} \in \Lambda_{\Gamma_n(C)}$ we have $f_{i,j}(A) = f_{i,j}(C)$. By Definition 3.11 for the general algorithm Γ_n it follows that $\Gamma_n(A) = \Gamma_n(C)$. But then $\Gamma_n(A) \cap B_{1/4}(1) \neq \emptyset$, which contradicts the assumption that $\Gamma_n(A) \subset \{0, 2\} + B_{2^{-n}}(0)$.

For the result $\{\Xi, \Omega_C\} \notin \Pi_1$ we argue similarly and suppose there is a sequence of algorithms Γ_n with $\mathrm{Sp}(A) \subset \Gamma_n(A) + B_{2^{-n}}(0)$ for all $A \in \Omega_C$. Define $A, C \in \Omega_C$ as above such that $\mathrm{Sp}(A) = \{0, 2\}$ and $\mathrm{Sp}(C) = \{0, 1\}$. By assumption there exists n such that $\Gamma_n(C) \cap B_{3/4}(2) = \emptyset$. Now choose $k > C(N, n)$ as above such that $\Gamma_n(A) = \Gamma_n(C)$. But then $\Gamma_n(A) \cap B_{3/4}(2) = \emptyset$, contradicting the assumption $2 \in \Gamma_n(A) + B_{2^{-n}}(0)$. △

Remark 3.22. The above arguments work equally well even if we restrict ourselves to self-adjoint compact operators (acting on any separable Hilbert space). Since no assumption was made on the model of computation used by the general algorithms, the result $\{\Xi, \Omega_C\} \notin \Sigma_1 \cup \Pi_1$ includes in particular the Turing model of computation, but also shows that a tower of height one is insufficient in any "reasonable" theory of computation (obeying Definition 3.11) in order to get a controlled convergence from above (Π_1) or from below (Σ_1). This suggests that more information than just matrix elements is required in order to build algorithms with Π_1 or Σ_1-error control. Indeed, it is shown in [5, Theorem 7.5] that in the compact case, a height two tower is sufficient for error control from below and above, while in the bounded self-adjoint case a height two tower only gives error control from below.

3.3 Spectral computations for unbounded operators

Having defined the appropriate SCI Hierarchy for computational spectral problems, we give the first basic results on how to apply this theory to unbounded linear operators. Later on, we will reduce the problem of computing the spectrum of a Dirac operator to the case we deal with in this section. First we establish some terminology and elementary results needed for the construction of the algorithm.

Injection modulus and resolvent norm

For a closed operator A, its *injection modulus* is defined as

$$\sigma_1(A) = \inf\{\|Ax\| \colon x \in \mathcal{D}(A), \|x\| = 1\}. \tag{3.9}$$

In the case that $\mathcal{D}(A)$ is finite dimensional, $\sigma_1(A)$ is simply the smallest singular value of A, i.e., the square root of the smallest eigenvalue A^*A. For any closed operator A, define the function

$$\gamma(z, A) \coloneqq \min\{\sigma_1(A - zI), \sigma_1(A^* - \bar{z}I)\},$$

and note that for an arbitrary vector $y \in \mathcal{D}(A)$ we have $\|Ay\| \geq \sigma_1(A)\|y\|$.

Recall the notation $R(z, A)$ for $(A - zI)^{-1}$ whenever $z \in \rho(A)$. One central observation is the relationship between the injection modulus and the reciprocal of the resolvent norm, which is explored by the two following lemmas from [8]:

Lemma 3.23. *Let A be a closed and densely defined operator. Then $\gamma(z, A) = \|R(z, A)\|^{-1}$ where we use the convention that $\|R(z, A)\|^{-1} = 0$ if $z \in \mathrm{Sp}(A)$, i.e., the resolvent norm is infinite.*

Proof. First assume that $z \notin \mathrm{Sp}(A)$. For an arbitrary $x \in \mathcal{D}(A)$ with $\|x\| = 1$ we have

$$1 = \|x\| = \|R(z, A)(A - zI)x\| \leq \|R(z, A)\| \, \|(A - zI)x\|,$$

and so $\sigma_1(A - zI) \geq \|R(z, A)\|^{-1}$. For the inequality in the opposite direction, pick $x_n \in \mathcal{D}(A)$ such that $\|x_n\| = 1$ and $\|R(z, A)x_n\| \to \|R(z, A)\|$. Then for all n,

$$1 = \|x_n\| = \|(A - zI)R(z, A)x_n\| \geq \sigma_1(A - zI) \, \|R(z, A)x_n\|,$$

i.e. $\sigma_1(A - zI) \leq \|R(z, A)x_n\|^{-1}$ and taking $n \to \infty$ we get $\sigma_1(A - zI) \leq \|R(z, A)\|^{-1}$. The same argument works for $\sigma_1(A^* - \bar{z}I)$ since $R(\bar{z}, A^*) = R(z, A)^*$ and $\|R(z, A)^*\| = \|R(z, A)\|$. Hence we have shown that $\gamma(z, A) = \|R(z, A)\|^{-1}$ if $z \in \mathrm{Sp}(A)$.

Now consider $z \in \mathrm{Sp}(A)$. We want to show that $\gamma(z, A) = 0$. If $\mathcal{N}(A - zI) \neq \{0\}$ then we are done, so assume that $A - zI$ is injective, and similarly assume $A^* - \bar{z}I$ is injective.

By definition, $(A - zI)^{-1}(A - zI)x = x$ for all $x \in \mathcal{D}(A)$. The range $\mathcal{R}(A - zI)$ is dense since $\mathcal{R}(A - zI)^{\perp} = \mathcal{N}((A - zI)^*) = \{0\}$. By injectivity, we can define the inverse $(A - zI)^{-1}$ on the dense domain $\mathcal{R}(A - zI)$ by the bounded map $(A - zI)x \mapsto x$. Then by the BLT-theorem, $(A - zI)^{-1}$ extends uniquely to a bounded operator defined everywhere, which we also denote by $(A - zI)^{-1}$. Since A, and hence $A - zI$, is closed we have $(A - zI)(A - zI)^{-1} = I$. To see this, pick an arbitrary $y \in \mathcal{R}(A - zI)$. Then $y = (A - zI)x$ for some $x \in \mathcal{D}(A)$ and by definition of $(A - zI)^{-1}$ we have

$$(A - zI)(A - zI)^{-1}y = (A - zI)(A - zI)^{-1}(A - zI)x = (A - zI)x = y$$

so $(A - zI)(A - zI)^{-1}$ is the identity on the dense subset $\mathcal{R}(A - zI)$ and so its unique bounded extension to the entire space must be I. Thus $(A - zI)^{-1}$ is a bounded inverse of $(A - zI)$ which means $z \notin \mathrm{Sp}(A)$, a contradiction. ■

Remark 3.24. If A is self-adjoint we have $\mathrm{Sp}(A) \subset \mathbb{R}$ so $(A - zI)^* = A^* - \bar{z}I = A - zI$ for $z \in \mathrm{Sp}(A)$. Thus by the proof of Lemma 3.23, $\|R(z, A)\|^{-1} = \sigma_1(A - zI)$.

Lemma 3.25. *Let A be a self-adjoint operator where the linear span of an orthonormal basis $\{e_n\}_{n=1}^{\infty}$ forms a core for A. Then the sequence of functions*

$$\gamma_n(z, A) \coloneqq \sigma_1((A - zI)P_n),$$

where P_n is the orthogonal projection onto $\mathrm{Span}\{e_1, \ldots, e_n\}$*, converges uniformly from above to*

$$\gamma(z, A) \coloneqq \sigma_1(A - zI) = \|R(z, A)\|^{-1}$$

on compact subsets of $\mathbb{C}$*.*

Proof. By the definition (3.9), as n increases we are taking the minimum over a strictly larger set in $\sigma_1((A - zI)P_n)$, so it is clearly non-increasing in n and no less than $\sigma_1(A - zI)$ for all n, where the minimum is taken over the entire space. Let $\varepsilon > 0$ be arbitrary and pick $x \in \mathcal{D}(A)$ with $\|x\| = 1$ such that $\|(A - zI)x\| \leq \sigma_1(A - zI) + \varepsilon$. Such an x must exists since if it did not then we would have the contradiction

$$\|(A - zI)x\| > \sigma_1(A - zI) + \varepsilon \text{ for all } x \in \mathcal{D}(A), \|x\| = 1.$$

Since $\mathcal{S} \coloneqq \mathrm{Span}\{e_n : n \in \mathbb{N}\}$ is a core for A, there is a sequence $x_n \in \mathcal{S}$ such that $x_n \to x$ and $Ax_n \to Ax$. Since each x_n is a finite linear combination of the basis vectors $\{e_n\}_n$, we can take a subsequence $x_{n_j} \in P_{n_j}(\mathcal{H}) = \mathrm{Span}\{e_1, \ldots, e_{n_j}\}$ so that $P_{n_j} x_{n_j} = x_{n_j}$. By adding a multiple of e_{n_j+1} if necessary, we may assume that $\|x_{n_j}\| = 1$. Then $P_{n_j} x_{n_j} \to x$ and $AP_{n_j} x_{n_j} \to Ax$ as $n_j \to \infty$. Then,

$$\begin{aligned}\sigma_1(A - zI) &\leq \sigma_1((A - zI)P_{n_j}) \\ &\leq \frac{\|(A - zI)P_{n_j} x_{n_j}\|}{\|P_{n_j} x_{n_j}\|} \to \|(A - zI)x\| \leq \sigma_1(A - zI) + \varepsilon.\end{aligned}$$

As $\varepsilon > 0$ was arbitrary, this shows that $\sigma_1((A - zI)P_n)$ converges to $\sigma_1(A - zI)$. Since the convergence is necessarily monotone from above, Dini's theorem [21, p. 85] implies that the functions $\gamma_n(z, A)$ converge uniformly to $\gamma(z, A)$ on compact subsets of $\mathbb{C}$. ■

Algorithm construction

We will now show how to construct an algorithm which outputs sets $\Gamma_n(A)$ converging to $\mathrm{Sp}(A)$ in the Attouch-Wets topology, for an unbounded closed operator A, following [8]. First, since the algorithm works by approximating the reciprocal resolvent norm, we need to impose some sort of control on how fast $\|R(z, A)\|^{-1}$ grows when z approaches the spectrum. If the growth is uncontrollable, then no Σ_1-result will be possible [5]. This motivates the following definition:

Definition 3.26. Let $g\colon [0, \infty) \to [0, \infty)$ be a continuous strictly increasing function, vanishing at 0 with $\lim_{x\to\infty} g(x) = \infty$ and $g(x) \leq x$. A closed operator A with non-empty spectrum has *controlled growth of the resolvent by* g if

$$\|R(z, A)\|^{-1} \geq g(\mathrm{dist}(z, \mathrm{Sp}(A))) \quad \forall z \in \mathbb{C}, \tag{3.10}$$

with the convention that $\|R(z, A)\|^{-1} = 0$ if $z \in \mathrm{Sp}(A)$. ▲

Note that this includes all self-adjoint (and more generally, normal) operators since we can take $g(x) = x$ in (3.10). Definition 3.26 provides a converse to the relation

$$\|R(z, A)\|^{-1} \leq \mathrm{dist}(z, \mathrm{Sp}(A)) \quad \forall z \in \mathbb{C}, \tag{3.11}$$

which holds for *all* closed operators by a standard Neumann series argument and shows why without an assumption like (3.10), the resolvent norm may blow up without control as z approaches $\mathrm{Sp}(A)$.

For each $n \in \mathbb{N}$, define the grid

$$\mathrm{Grid}(n) \coloneqq \frac{1}{n}(\mathbb{Z} + i\mathbb{Z}) \cap B_n(0),$$

where $B_n(0)$ is the closed ball of radius n around 0 in $\mathbb{C}$. Then $\mathrm{Grid}(n)$ is a finite subset of $B_n(0)$ with spacing $1/n$ between its points. Given the continuous function $g\colon [0,\infty) \to [0,\infty)$, which is strictly increasing with $g(0) = 0$ and diverging at infinity, let h denote the inverse function $h(y) = g^{-1}(y)$ for $y \in [0,\infty)$. Then $h(0) = 0$, h is strictly increasing and has $h(y) \geq y$ since h is the reflection of g across the line $y = x$. For each $n \in \mathbb{N}$ define the function $h_n\colon [0,\infty) \to [0,\infty)$ by

$$h_n(y) \coloneqq \min_{k\in\mathbb{N}}\{k/n\colon g(k/n) > y\} = \min_{k\in\mathbb{N}}\{k/n\colon k/n > h(y)\}. \tag{3.12}$$

Note that h_n can be computed on any input $y \in \mathbb{Q}$ by a finite number of evaluations of g and that

$$h(y) \leq h_n(y) \leq h(y) + 1/n. \tag{3.13}$$

With this setup, we can formulate the steps of the algorithm for operators A as in Definition 3.26. As a fundamental assumption, we have a sequence of functions $\gamma_n(z, A)$ that converges uniformly to

$$\gamma(z, A) = \min\{\sigma_1(A - zI), \sigma_1(A^* - \bar{z}I)\} = \|R(z, A)\|^{-1}$$

on compact subsets of $\mathbb{C}$. If the functions $\gamma_n(z, A)$ can be computed to arbitrary precision via finite arithmetic means from the evaluation functions in Λ, the map $A \to \Gamma_n(A)$ described in Algorithm 1 defines an arithmetic algorithm in the sense of Definition 3.11 for each n [8, pp. 20–21].

Algorithm 1: A general algorithm to estimate the spectrum.

Input: A and g as in Definition 3.26, a sequence of functions $\gamma_n(z, A)$ converging uniformly to $\gamma(z, A)$ on compact subsets of $\mathbb{C}$, $h_n\colon [0,\infty) \to [0,\infty)$ as in (3.12), $n \in \mathbb{N}$.
Output: Approximation $\Gamma_n(A) \subset \mathbb{C}$ to $\mathrm{Sp}(A)$.
for $z \in \mathrm{Grid}(n)$ **do**
  if $\gamma_n(z, A) \leq (|z|^2 + 1)^{-1}$ **then**
    $r_{n,z} \coloneqq h_n(\gamma_n(z, A))$;
    $\Upsilon_{n,z} \coloneqq B_{r_{n,z}}(z) \cap \mathrm{Grid}(n)$;
    $M_z^n \coloneqq \{w \in \Upsilon_{n,z}\colon \gamma_n(w, A) = \min_{v\in\Upsilon_{n,z}} \gamma_n(v, A)\}$
  else
    $M_z^n \coloneqq \emptyset$
  end
end
$\Gamma_n(A) \coloneqq \cup_{z\in\mathrm{Grid}(n)} M_z^n$

Proposition 3.27. *Let $\{\Gamma_n\}_{n=1}^{\infty}$ be the height one arithmetic tower for computing the problem function $\Xi(A) = \mathrm{Sp}(A)$ where $\Gamma_n(A)$ is defined in Algorithm 1. By taking a subsequence if necessary, we can assume that each set Γ_n is non-empty.*

Proof. By construction, $\Gamma_n(A)$ is empty if and only if $\gamma_n(z, A) > (|z|^2 + 1)^{-1}$ for all $z \in \mathrm{Grid}(n)$. Assuming $\mathrm{Sp}(A) \neq \emptyset$, we can choose m large so that $B_m(0)$ intersects $\mathrm{Sp}(A)$. Then there exists $z \in B_m(0)$ with $\mathrm{dist}(z, \mathrm{Sp}(A)) < \frac{1}{2}((m+1)^2 + 1)^{-1}$. By (3.11), we have $\gamma(z, A) < \frac{1}{2}(m^2 + 1)^{-1}$. By taking n large enough, we can choose a point $z_n \in \mathrm{Grid}(n)$ arbitrarily close to z. We can safely assume that $|z_n - z| < 1$, and by continuity of γ, we can in particular choose this z_n so that $|\gamma(z_n, A) - \gamma(z, A)| < \frac{1}{2}((m+1)^2 + 1)^{-1}$. Thus

$$\gamma(z_n, A) < \gamma(z, A) + \frac{1}{2}((m+1)^2 + 1)^{-1} < ((m+1)^2 + 1)^{-1} < (|z_n|^2 + 1)^{-1},$$

where the last inequality holds because $|z_n| \leq |z| + |z - z_n| < m + 1$. Since $\gamma_n \to \gamma$ locally uniformly, this also shows that for large enough n, there exists $z_n \in \mathrm{Grid}(n)$ such that $\gamma_n(z_n, A) \leq (|z_n|^2 + 1)^{-1}$, i.e., $\Gamma_n(A)$ is non-empty. ■

The next step is to show that the sets $\Gamma_n(A)$ defined in Algorithm 1 do converge to $\mathrm{Sp}(A)$ in the Attouch-Wets topology, and later to show how to get Σ_1-error control. For the Attouch-Wets convergence, we give a slightly simplified version of [8, Proposition 6.5] clarifying some details of the proof.

Proposition 3.28. *Suppose A is a closed operator with non-empty spectrum and that we have a sequence of functions $\gamma_n(z, A)$ that converge uniformly to $\gamma(z, A) = \|R(z, A)\|^{-1}$ on compact subsets of $\mathbb{C}$. Suppose also that A has controlled growth of the resolvent by g, i.e.*

$$g(\mathrm{dist}(z, \mathrm{Sp}(A))) \leq \|R(z, A)\|^{-1} \qquad \forall z \in \mathbb{C}. \tag{3.14}$$

Then the sets $\Gamma_n(A)$ as defined in Algorithm 1 converge to $\mathrm{Sp}(A)$ in the Attouch-Wets topology.

Proof. As explained above, we may assume without loss of generality that $\Gamma_n(A)$ is non-empty for every n. Note also that (3.14) is equivalent to

$$\mathrm{dist}(z, \mathrm{Sp}(A)) \leq h(\gamma(z, A)) \qquad \forall z \in \mathbb{C}, \tag{3.15}$$

where $h = g^{-1}$ as in (3.12). We will use this version several times in the proof. To show the convergence we use the characterisation of the Attouch-Wets topology given earlier: $d_{\mathrm{AW}}(C_n, C) \to 0$ if and only if for any $\delta > 0$ and compact K there is a finite $N \in \mathbb{N}$ such that if $n > N$ then $C_n \cap K \subset C + B_\delta(0)$ and $C \cap K \subset C_n + B_\delta(0)$. By fixing an $m \in \mathbb{N}$ sufficiently large that $B_m(0) \cap \mathrm{Sp}(A) \neq \emptyset$, it is sufficient to show that given $\delta > 0$, there exists N such that if $n > N$ then the inclusions

$$\mathrm{Sp}(A) \cap B_m(0) \subset \Gamma_n(A) + B_\delta(0), \tag{3.16}$$

$$\Gamma_n(A) \cap B_m(0) \subset \mathrm{Sp}(A) + B_\delta(0) \tag{3.17}$$

both hold. We fix such an m for the rest of the proof and introduce the notation $\varepsilon_n = \|\gamma_n(\cdot, A) - \gamma(\cdot, A)\|_{\infty, B_{m+1}(0)}$ to mean the supremum norm over the closed

ball $B_{m+1}(0)$. By assumption, $\varepsilon_n \to 0$ and by passing to a subsequence of $\{\gamma_n\}_{n=1}^{\infty}$ if necessary, we may assume that $\varepsilon_{n+1} \leq \varepsilon_n$ for all n.

First we show the inclusion (3.16). Suppose that $z \in \mathrm{Sp}(A) \cap B_m(0)$. For any $n \geq m$ there is a $w \in \mathrm{Grid}(n)$ with $|w - z| < 1/n$. Then

$$\gamma_n(w, A) \leq \gamma(w, A) + \varepsilon_n \leq \mathrm{dist}(w, \mathrm{Sp}(A)) + \varepsilon_n \leq 1/n + \varepsilon_n. \tag{3.18}$$

Clearly there exists $N \in \mathbb{N}$ such that for all $n > N$ we have $\varepsilon_n, 1/n \leq \frac{1}{2}((m+1)^2 + 1)^{-1}$ and putting this into (3.18), $\gamma_n(w, A) < (|w|^2 + 1)^{-1}$ since $|w| \leq |z| + 1/n \leq m + 1/n < m + 1$. Thus $M_w^n \neq \emptyset$ for $n > N$. Now pick some arbitrary $y \in M_w^n$. By definition of M_w^n as a subset of $\Upsilon_{n,w}$ in Algorithm 1, we have (recall (3.13))

$$|y - w| \leq r_{n,w} = h_n(\gamma_n(w, A)) \leq h(\gamma_n(w, A)) + 1/n,$$

and so

$$|y - z| \leq |w - z| + |y - w| \leq 1/n + 1/n + h(\gamma_n(w, A)).$$

Since $\gamma_n(w, A) \leq \varepsilon_n + 1/n$ and h is strictly increasing,

$$|y - z| \leq 2/n + h(\varepsilon_n + 1/n). \tag{3.19}$$

Recall that $\lim_{x \to 0} h(x) = 0$. Thus in the above choice of N, we may also choose N such that for $n > N$ we have $2/n + h(\varepsilon_n + 1/n) \leq \delta$. Hence by taking n sufficiently large in the reasoning above we can choose $w \in \mathrm{Grid}(n)$ and corresponding $y \in M_w^w$, so that by (3.19) we have

$$|y - z| \leq \delta.$$

Of course $y \in M_w^n \subset \Gamma_n(A)$ and since $z \in \mathrm{Sp}(A) \cap B_m(0)$ was arbitrary, this shows the inclusion $\mathrm{Sp}(A) \cap B_m(0) \subset \Gamma_n(A) + B_\delta(0)$ for all $n > N$.

For the second inclusion, suppose for contradiction that it does not hold for all n greater than some finite threshold. That is, for a given $\delta > 0$, suppose there is a subsequence $\{\Gamma_{n_j}\}$ of $\{\Gamma_n\}$ such that for each n_j there is a point $z_{n_j} \in \Gamma_{n_j}(A) \cap B_m(0)$ with $\mathrm{dist}(z_{n_j}, \mathrm{Sp}(A)) > \delta$. By definition, $z_{n_j} \in M_{w_{n_j}}^{n_j}$ for some $w_{n_j} \in \mathrm{Grid}(n_j)$. Let $y_{n_j} \in \mathrm{Sp}(A)$ be of minimal distance from $w_{n_j} \in B_{n_j}(0)$. That such a y_{n_j} exists for each w_{n_j}, follows from the extreme value theorem after considering the continuous function $z \mapsto |z - w_{n_j}|$ on the compact set $\mathrm{Sp}(A) \cap B_{m(n_j)}(0)$ where $m(n_j) \in \mathbb{N}$ is chosen to be large enough that $\mathrm{Sp}(A) \cap B_{m(n_j)}(0)$ is non-empty. By the choice of y_{n_j} and (3.15),

$$|y_{n_j} - w_{n_j}| = \mathrm{dist}(w_{n_j}, \mathrm{Sp}(A)) \leq h(\gamma(w_{n_j}, A)).$$

Next, the idea is to choose a point $v_{n_j} \in I(j) \coloneqq \Upsilon_{n_j, w_{n_j}}$ close to y_{n_j} (see Section 3.3). Now $I(j)$ is exactly the set of points in $\mathrm{Grid}(n_j)$ that are at most $r_{n_j} \coloneqq h_{n_j}(\gamma_{n_j}(w_{n_j}, A))$ away from w_{n_j}. Start at the point w_{n_j} which is clearly in $I(j)$. Then move in a straight line towards y_{n_j} for a distance of exactly r_{n_j} and label the endpoint of this movement u_{n_j}. Since u_{n_j} is on the straight line between w_{n_j} and y_{n_j},

$$\begin{aligned} |u_{n_j} - y_{n_j}| &= |w_{n_j} - y_{n_j}| - |w_{n_j} - u_{n_j}| \\ &\leq h(\gamma(w_{n_j}, A)) - h_{n_j}(\gamma_{n_j}(w_{n_j}, A)) \\ &\leq h(\gamma(w_{n_j}, A)) - h(\gamma_{n_j}(w_{n_j}, A)). \end{aligned}$$

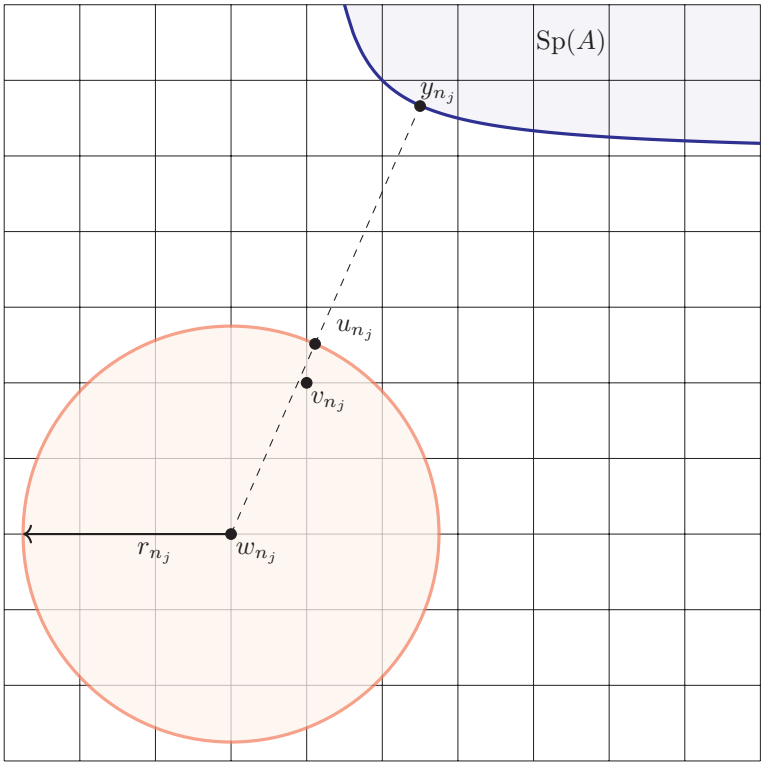


Figure 3.1: Grid(n_j) with Sp(A) shaded in blue. $I(j)$ is exactly the vertices of Grid(n_j) within a radius $r_{n_j} \coloneqq h_{n_j}(\gamma_{n_j}(w_{n_j}, A))$ from w_{n_j}.

Now u_{n_j} may not be exactly on Grid(n_j) and hence not in $I(j)$. The "worst case" scenario is that in order to get from u_{n_j} to a point in $I(j)$, we have to move a length $1/n_j$ away from y_{n_j} along both the real and imaginary axes, thus adding a total distance that is less than $2/n_j$. Hence, given the point u_{n_j} we can find a point $v_{n_j} \in I(j)$ such that

$$|v_{n_j} - y_{n_j}| \leq \frac{2}{n_j} + h(\gamma(w_{n_j}, A)) - h(\gamma_{n_j}(w_{n_j}, A)). \tag{3.20}$$

Recall that z_{n_j} is chosen from $I(j) = \Upsilon_{n_j, w_{n_j}}$ so that z_{n_j} minimizes $\gamma_{n_j}(\cdot, A)$ over $I(j)$, which in particular contains w_{n_j} and v_{n_j}. We know that $\gamma_{n_j}(w_{n_j}, A) < (|w_{n_j}|^2 + 1)^{-1}$ since $M^{n_j}_{w_{n_j}}$ is non-empty by assumption. Then it follows that

$$\gamma(z_{n_j}, A) \leq \gamma_{n_j}(z_{n_j}, A) + \varepsilon_{n_j} \leq \min\left\{\frac{1}{|w_{n_j}|^2 + 1}, \gamma_{n_j}(v_{n_j}, A)\right\} + \varepsilon_{n_j}.$$

Recalling (3.15) and applying the increasing function h to the above inequality gives

$$0 < \delta \leq \operatorname{dist}(z_{n_j}, \operatorname{Sp}(A)) \leq h\left(\min\left\{\frac{1}{|w_{n_j}|^2 + 1}, \gamma_{n_j}(v_{n_j}, A)\right\} + \varepsilon_{n_j}\right). \tag{3.21}$$

Since $\delta > 0$, the right hand side of (3.21) cannot become arbitrarily close to zero. Now $\lim_{x \to 0} h(x) = 0$ and $\varepsilon_{n_j} \to 0$, so in particular $\sup_{n_j} |w_{n_j}| < \infty$, i.e.,

the sequence $\{w_{n_j}\}$ is contained in a bounded subset of $\mathbb{C}$, and thus so is $\{v_{n_j}\}$. The local uniform convergence $\gamma_{n_j} \to \gamma$ and continuity of h together imply

$$\frac{2}{n_j} + h(\gamma(w_{n_j}, A)) - h(\gamma_{n_j}(w_{n_j}, A)) \to 0, \qquad \text{as } n_j \to \infty.$$

But then by (3.20) (also recall (3.11) for the first inequality),

$$\gamma(v_{n_j}, A) \leq \operatorname{dist}(v_{n_j}, \operatorname{Sp}(A)) \leq |v_{n_j} - y_{n_j}| \to 0.$$

Again the local uniform convergence gives $\gamma_{n_j}(v_{n_j}, A) \to 0$ as well, contradicting (3.21). Thus the existence of the sequence $\{z_{n_j}\}$ as defined is impossible, which shows that given $\delta > 0$, for all large enough n we have the inclusion $\Gamma_n(A) \cap B_m(0) \subset \operatorname{Sp}(A) + B_\delta(0)$, finishing the proof. ■

Proposition 3.28 establishes that for an operator A satisfying the assumptions in Algorithm 1, the sets $\Gamma_n(A)$ do indeed converge to $\operatorname{Sp}(A)$ in the Attouch-Wets topology. It is perhaps unsurprising that with knowledge of functions $\gamma_n(z, H)$ which converge to $\gamma(z, H) = \|R(z, H)\|^{-1}$ one can construct an iterative algorithm that searches a grid of points and converges to the spectrum in for example the Attouch-Wets metric. The key feature of Algorithm 1 however, is that under quite mild assumptions on the convergence of the γ_n this construction allows us to compute a rigorous bound on the error of the approximation, which is the heart of the Σ_1-class.

In order to achieve convergence with Σ_1-error control as in Definition 3.18, we use the criterion given by Lemma 3.19. According to this result, if $\Gamma_n(A) \to \operatorname{Sp}(A)$ in the Attouch-Wets topology for each $A \in \Omega$, then in order to prove Σ_1-convergence, it is sufficient to show the existence of functions $E_n \colon \mathbb{C} \to [0, \infty)$ converging uniformly to zero on compact subsets of $\mathbb{C}$, such that E_n can be computed over $\Gamma_n(A)$, and

$$\operatorname{dist}(z, \operatorname{Sp}(A)) \leq E_n(z) \qquad \forall z \in \Gamma_n(A).$$

This is what we show next (adapted from [8, p. 24]). Let $\Omega_g(\mathcal{H})$ denote the domain consisting of all closed operators on a given separable Hilbert space $\mathcal{H}$ which have growth of the resolvent bounded by a function g as in Definition 3.26.

Proposition 3.29. *Let $\mathcal{H}$ be a separable Hilbert space and $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ the computational problem with $\Omega = \Omega_g(\mathcal{H})$ and $\Xi(A) = \operatorname{Sp}(A)$ for every $A \in \Omega$. $(\mathcal{M}, d)$ is the space of closed subsets of $\mathbb{C}$ with the Attouch-Wets metric and Λ contains the functions $A \mapsto g(i/j)$ for all $i, j \in \mathbb{N}$ as well as information sufficient to compute arithmetic functions $\gamma_n(z, A)$ that converge uniformly **from above** to $\gamma(z, A)$ on compact subsets of $\mathbb{C}$ for each $A \in \Omega$.*

Then the sequence Γ_n of general algorithms defined in Algorithm 1 define a height one tower of arithmetic algorithms for $\{\Xi, \Omega, \mathcal{M}, \Lambda\}$ with error control as described by the class Σ_1 in Definition 3.18.

Proof. By the construction in Algorithm 1 and the definition of Λ, we have an arithmetic tower of height one, since Proposition 3.28 proves convergence in the Attouch-Wets metric. We are left with showing that the assumptions for Lemma 3.19 are satisfied in order to get the error control.

Let $A \in \Omega$ be arbitrary and recall that we may assume $\Gamma_n(A)$ is always non-empty. Set

$$E_n(z) \coloneqq h_n(\gamma_n(z, A))$$

for $z \in \Gamma_n(A)$ and zero on $\mathbb{C} \setminus \Gamma_n(A)$, with h_n as in (3.12). By our assumptions on Λ, the functions E_n are computable from the available information. Since we have $\|R(z, A)\|^{-1} = \gamma(z, A)$ and we assume $\gamma_n \to \gamma$ locally uniformly and from above, the assumption

$$g(\operatorname{dist}(z, \operatorname{Sp}(A))) \leq \|R(z, A)\|^{-1}$$

gives

$$\operatorname{dist}(z, \operatorname{Sp}(A)) \leq h(\gamma(z, A)) \leq E_n(z),$$

for all $z \in \Gamma_n(A)$. Now suppose for contradiction that E_n does not converge uniformly to zero on compact subsets of $\mathbb{C}$. Then there exists a compact set K, an $\varepsilon > 0$ and a sequence $\{z_{n_j}\}$ in K such that $E(z_{n_j}) \geq \varepsilon$ for all n_j. By definition of E_{n_j} this means $z_{n_j} \in \Gamma_{n_j}$. Since K is compact, we may assume (by passing to a subsequence if needed) that $z_{n_j} \to z$ for some $z \in \mathbb{C}$. Since $\Gamma_{n_j}(A) \to \operatorname{Sp}(A)$ in the Attouch-Wets metric, the points z_{n_j} become arbitrarily close to $\operatorname{Sp}(A)$ and so their limit z must be in $\operatorname{Sp}(A)$. Hence by the convergence of γ_{n_j} we must have $\gamma_{n_j}(z_{n_j}, A) \to \gamma(z, A) = 0$. Then because $h \leq h_n \leq h + 1/n$ and $\lim_{y \to 0} h(y) = 0$, we must have

$$E_{n_j}(z_{n_j}) = h_{n_j}(\gamma_{n_j}(z_{n_j}, A)) \leq h(\gamma_{n_j}(z_{n_j}, A)) + 1/n_j \to 0$$

as $n_j \to \infty$, which contradicts the assumption $E(z_{n_j}) \geq \varepsilon$ for all n_j. Thus E_n must converge uniformly to zero on compact subsets of $\mathbb{C}$. In particular for all $m \in \mathbb{N}$ this gives

$$\lim_{n \to \infty} \sup_{z \in \Gamma_n(A) \cap B_m(0)} E_n(z) = 0,$$

and so the conditions of Lemma 3.19 have all been satisfied. ■

Finally, we remark that if the computational domain consists of exclusively self-adjoint operators, then we can obviously simplify Algorithm 1 by only considering grids of real numbers, and all the results in this section still hold. This is what one would do in practice, but even though we will deal with self-adjoint Dirac-operators in the next chapters, we will not make this distinction going forward, since such a simplification does not affect the theoretical considerations in a material way.

CHAPTER 4

Numerical Integration

In order to construct the algorithms used to approximate spectra, certain inner products (i.e. integrals) must be computed to a given level of precision and to achieve this, some numerical methods will be required. This short chapter gathers the relevant terminology and results, whose proofs can be found in Chapters 2 and 3 of [23]. Section 4.1 review quasirandom sampling and Section 4.2 gives the results we will need on numerical integration. Code for producing the figures in this chapter may be found at *https://github.com/emilhaugen/qmc*.

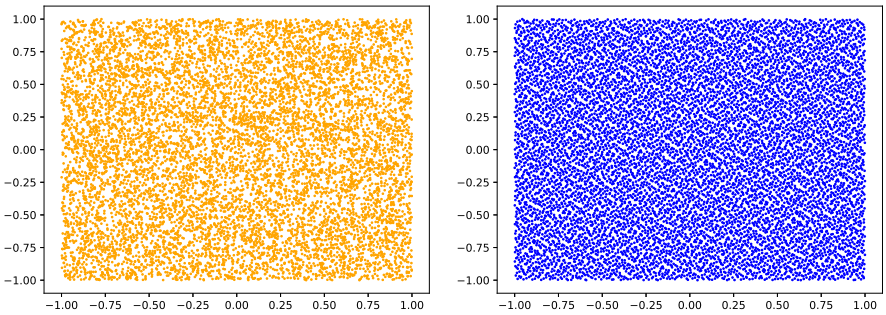
4.1 Quasirandom sampling

As the reader probably knows, Monte Carlo integration is a method for approximating an integral of a function over some domain in $\mathbb{R}^n$ by sampling points from the domain using a uniform probability distribution and taking the average value of the function over those sample points. This is widely used in high dimensional problems where analytical integration is often not an option. This gives error bounds based on probability, and is fundamentally a random algorithm. Since we want a deterministic algorithm, we instead sample over a "quasirandom" set, and the resulting estimation procedure is therefore called quasi-Monte Carlo integration. Another advantage is that whereas random samples can produce clusters of points in one area while missing other areas of the domain, quasirandom sequences seek to avoid this (see Figure 4.1). We begin with the one-dimensional case since this forms the basis for later generalisation to higher dimensions.

Quasirandom sequences

Informally, the "discrepancy" of a set of points is a measure of how much the points deviate from being uniformly distributed. The discrepancy will be high if the set contains dense clusters or if it leaves large empty spaces. A sample from the uniform distribution will *on average* have low discrepancy, but any specific instance may have very high discrepancy.

Given a positive integer, b, the *van der Corput sequence in base* b is a low-discrepancy sequence over the interval $[0, 1]$. For a given base b and $n \in \mathbb{N}$,



(a) Random Uniform (b) $(2, 3)$ Halton sequence

Figure 4.1: Sample of 10000 points.

we write the base b-expansion of n as

$$n = \sum_{k=0}^{L-1} d_k(n) b^k,$$

where $d_k(n)$ is the k-th digit of n in base b. The n-th number of the van der Corput sequence in base b, $v_b(n)$, is given by considering this expansion and reversing the digits by taking the sum

$$v_b(n) \coloneqq \sum_{k=0}^{L-1} d_k(n) b^{-(k+1)}. \tag{4.1}$$

For example, in base $b = 2$ the 13-th number of the van der Corput sequence is found by considering the binary representation of 13 which is 1101. Then (4.1) gives $v_2(13) = 1 \cdot 2^{-1} + 0 \cdot 2^{-2} + 1 \cdot 2^{-3} + 1 \cdot 2^{-4} = 11/16 = 0.6875$.

The *Halton sequence* generalises the van der Corput sequence to produce a sample of points in $[0, 1]^d$ for arbitrary dimension d. This is done by taking a van der Corput in base q_1 for the first coordinate, base q_2 for the second coordinate up to base q_d for the last coordinate where $q_1, \ldots, q_d$ are pairwise relatively prime integers. Given $a, b \in \mathbb{R}$ with $a < b$, letting t_n denote the n-th Halton point in dimension d, the rescaling $t_n \mapsto (a - b)t_n + (a, \ldots, a)^T =: s_n$ gives the *rescaled Halton points* in $[a, b]^d$.

In addition to replacing a probabilistic algorithm with a deterministic one, low-discrepancy sequences are often preferable because they can give improved numerical efficiency compared to standard uniform random sampling, due to the fact that they avoid the clustering found in random sequences. See Figure 4.2 and Figure 4.3 for examples (note that in one dimension, Halton sampling simply corresponds to the base 2 van der Corput sequence).

4.2 Some results from quasi-Monte Carlo integration

Next we recount some theoretical results from quasi-Monte Carlo integration, referring to [23, Chapters 2, 3] for details and proofs. These results will provide

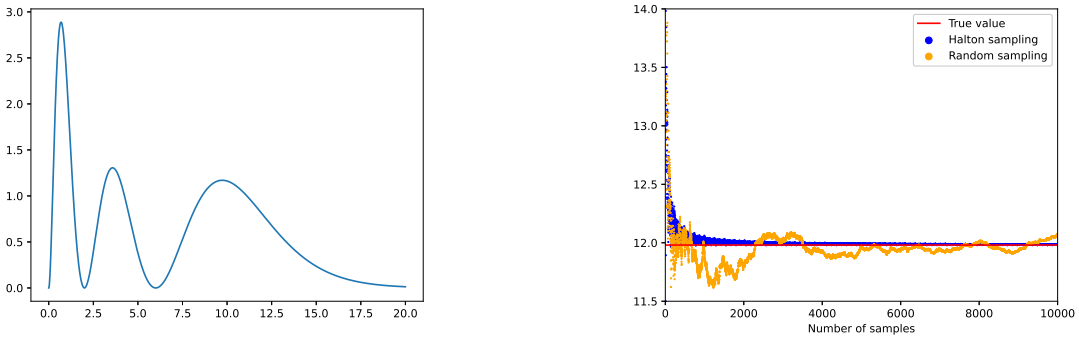


Figure 4.2: Graph of $f(x) = \frac{1}{4} \cdot (x^3 - 8x^2 + 12x)^2 \cdot \exp(-x)$ (left) and estimated integrals on $[0, 20]$.

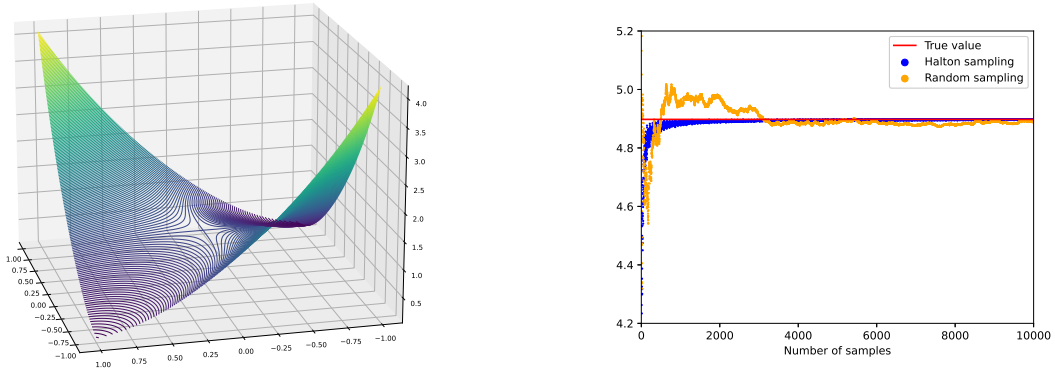


Figure 4.3: Graph of $f(x, y) = \exp(-x^2 - y^2) + (x + y)^2$ (left) and estimated integrals over $[-1, 1]^2$.

error bounds which will be important later. First we need a formal definition of discrepancy.

Star discrepancy

Let $\{t_1, \dots t_j\}$ be a set of points in $[0, 1]^d$. For an arbitrary subset $B \subset [0, 1]^d$, we define the number

$$P(B; \{t_k\}) \coloneqq \frac{1}{j} \sum_{k=1}^{j} \chi_B(t_k),$$

where χ_B is the characteristic function of B. In other words, $P(B; \{t_k\})$ is the function which for each B counts the proportion of elements from $\{t_1, \dots t_j\}$ which are contained in B.

Definition 4.1. (Star discrepancy.) Let $\{t_1, \dots t_j\}$ be a sequence in $[0,1]^d$ and let $\mathcal{K}$ denote all subsets of $[0,1]^d$ on the form $\prod_{k=1}^{d}[0, y_k)$ for $y_k \in (0,1]$. The *star discrepancy* of $\{t_1, \dots t_j\}$ is defined by

$$D_j^*(\{t_1, \dots t_j\}) = \sup_{K \in \mathcal{K}} |P(K; \{t_k\}) - \lambda(K)|,$$

where $\lambda(K)$ is the Lebesgue-measure of K. ▲

Let $\{t_1, \dots t_j\}$ and $K = \prod_{k=1}^{d}[0, y_k)$ be as in Definition 4.1. Note that we will always have $0 < \lambda(K) \leq 1$ and that K is star-shaped with respect to the origin (hence the name). Sets on the form of K are called *subintervals* of $[0,1]^d$. The quantity $|P(K; \{t_k\}) - \lambda(K)|$ will be small if the proportion of points $t_k \in K$ is close to $\lambda(K)$ and it will be large if the proportion of points $t_k \in K$ is very different from $\lambda(K)$. Since $\lambda(K)$ is simply the proportion of the total space $[0,1]^d$ which is occupied by K, we see that the star-discrepancy does indeed measure how close the sequence is to being equidistributed, as claimed earlier. In the case of the Halton sequence, we have the following fundamental result [23, p. 29]:

Theorem 4.2. *If $\{t_k\}_{k \in \mathbb{N}}$ is the Halton sequence in $[0,1]^d$ in the pairwise relatively prime bases $q_1, \dots, q_d$ then*

$$D_j^*(\{t_1, \dots t_j\}) < \frac{d}{j} + \frac{1}{j} \prod_{k=1}^{d} \left(\frac{q_k - 1}{2 \log(q_k)} \log(j) + \frac{q_k + 1}{2} \right).$$

Given the dimension d and bases $q_1, \dots, q_d$ one can easily compute in finitely many arithmetic operations and comparisons a constant $C(d)$ which only depends on d such that Theorem 4.2 implies

$$D_j^*(\{t_1, \dots t_j\}) < C(d) \frac{(\log(j) + 1)^d}{j}. \tag{4.2}$$

To see this, note that we can safely assume that $q_1 < \cdots < q_d$ and so we have,

$$\begin{aligned} \frac{d}{j} + \frac{1}{j} \prod_{k=1}^{d} \left(\frac{q_k - 1}{2 \log(q_k)} \log(j) + \frac{q_k + 1}{2} \right) &< \frac{d}{j} + \frac{1}{j} \left(q_d \log(j) + q_d \right)^d \\ &< (d + q_d^d) \frac{(\log(j) + 1)^d}{j}. \end{aligned}$$

Thus we may for instance take $C(d) = d + q_d^d$ in (4.2). This is clearly not necessarily optimal, but will be sufficient for our purposes.

Error bounds

Before giving the main result, we need a few new definitions concerning the variation of a function, starting in the one dimensional case. Recall that for a function $f : [0,1] \to \mathbb{R}$, its *total variation* is defined by

$$V(f) \coloneqq \sup_{P \in \mathcal{P}} \sum_{i=0}^{n_p - 1} |f(x_{i+1}) - f(x_i)|$$

where the supremum is taken over all partitions of the interval, i.e.,

$$\mathcal{P} = \{P = \{x_0, \dots, x_{n_P}\} : P \text{ is a partition of } [0,1] \text{ with } x_i < x_{i+1}\}.$$

A function f is said to be of *bounded variation* if $V(f)$ is finite. Of course functions of bounded variation are in particular bounded and so they are integrable on $[0,1]$. A classical result called Koksma's inequality [23, p. 18] says that if f has bounded variation $V(f)$ on $[0,1]$ then for any $t_1, \dots, t_j \in [0,1]$,

$$\left| \frac{1}{j} \sum_{k=1}^{j} f(t_k) - \int_0^1 f(u)\mathrm{d}u \right| \leq V(f) D_j^*(\{t_1, \dots t_j\}).$$

The next goal is to give an analogous result in higher dimensions, which combined with (4.2) would guarantee that if we choose a high enough sample size j, then we can approximate the integral over $[0,1]^d$ to arbitrary precision using Halton sampling. For a function f on $[0,1]^d$ and a subinterval $J \subset [0,1]^d$, we denote by $\Delta(f; J)$ an alternating sum of values of f at the vertices of J so that function values at adjacent vertices (only differing in one coordinate) have opposite signs in the sum. The *Vitali variation* is defined by

$$V^{(d)}(f) \coloneqq \sup_{\mathcal{P}} \sum_{J \in \mathcal{P}} |\Delta(f; J)|,$$

where the supremum is taken over the collection $\mathcal{P}$ of all partitions of $[0,1]^d$ into subintervals. A simpler formula

$$V^{(d)}(f) = \int_0^1 \cdots \int_0^1 \left| \frac{\partial^d f}{\partial u_1 \cdots \partial u_d} \right| \mathrm{d}u_1 \cdots \mathrm{d}u_d \tag{4.3}$$

holds when the partial derivatives are continuous on $[0,1]^d$. For $k = 1, \dots, d$ and $1 \leq i_1 < i_2 < \cdots < i_k \leq d$ let $V^{(k)}(f; i_1, \dots i_k)$ be the Vitali variation of the restriction of f to the k-dimensional face $\{(u_1, \cdots u_d) \in [0,1]^d : u_j = 1 \text{ for } j \neq i_1, \dots, i_k\}$. For $k = 1, \dots, d$, let $\mathcal{I}_k$ denote the set of all strictly increasing multi-indices with length k, i.e.,

$$\mathcal{I}_k \coloneqq \{I = (i_1, \dots, i_k) : 1 \leq i_1 < i_2 < \cdots < i_k \leq d\}.$$

The *Hardy-Krause variation* of f is defined as

$$\mathrm{TV}_{[0,1]^d}(f) \coloneqq \sum_{k=1}^{d} \sum_{I \in \mathcal{I}_k} V^{(k)}(f; i_1, \dots, i_k). \tag{4.4}$$

By replacing the interval $[0,1]^d$ with a more general interval $[a,b]^d$ in the preceding discussion, we get analogous definitions of variation for functions defined on $[a,b]^d$. Note that if the partial derivatives exist as in (4.3) then we have the more computationally friendly expression for the Hardy-Krause variation,

$$\mathrm{TV}_{[a,b]^d}(f) = \sum_{k=1}^{d} \sum_{I \in \mathcal{I}_k} \int_a^b \cdots \int_a^b \left| \frac{\partial^k f}{\partial u_{i_1} \cdots \partial u_{i_k}}(\tilde{u}) \right| \mathrm{d}u_{i_1} \cdots \mathrm{d}u_{i_k}, \tag{4.5}$$

where $\tilde{u}_j = u_j$ for $j = i_1, \ldots i_k$ and $\tilde{u}_j = b$ otherwise. Of course, exact computation of (4.5) may be infeasible, but it can be used to establish upper bounds for the total variation which will usually be sufficient. With definitions established, we can state the main result in this chapter, the so-called Koksma-Hlawka inequality —see [23, p. 20] for the proof.

Theorem 4.3. *If f has bounded variation $\mathrm{TV}_{[0,1]^d}(f)$ on the cube $[0,1]^d$ then for any $\{t_1, \ldots t_j\} \subset [0,1]^d$,*

$$\left| \frac{1}{j} \sum_{k=1}^{j} f(t_k) - \int_{[0,1]^d} f(u)\,\mathrm{d}u \right| \leq \mathrm{TV}_{[0,1]^d}(f) D_j^*(\{t_1, \ldots t_j\}).$$

By re-scaling, if f has bounded variation $\mathrm{TV}_{[a,b]^d}(f)$ on the cube $[a,b]^d$ and $s_k = (b-a)t_k + (a, a, \ldots, a)^T$ are the re-scaled Halton points, then

$$\left| \frac{(b-a)^d}{j} \sum_{k=1}^{j} f(s_k) - \int_{[a,b]^d} f(u)\,\mathrm{d}u \right| \leq (b-a)^d \cdot \mathrm{TV}_{[a,b]^d}(f) D_j^*(\{t_1, \ldots t_j\}).$$

Thus we have established the main result on quasi-Monte Carlo integration which will be used later on. The key observation is the combination of Theorem 4.3 and Equation (4.2). For suppose $\delta > 0$, and that we have an upper bound M on the Hardy-Krause variation of the function f. Assuming logarithms can be computed to arbitrary precision (say, using a power series with Lagrange bound on the error), by successive computation, we can in finite time find j so that

$$(b-a)^d \cdot M \cdot C(d) \frac{(\log(j)+1)^d}{j} < \delta.$$

Then the Koksma-Hlawka inequality guarantees that $\frac{(b-a)^d}{j} \sum_{k=1}^{j} f(s_k)$ estimates the integral of f over $[a,b]^d$ with an error less than δ.

CHAPTER 5

Dirac Operator with Bounded Potential

This chapter concerns the computational spectral problem for the three-dimensional Dirac operator with bounded potential. In Section 5.1 we introduce the operator, its domain of self-adjointness and formulate the computational problem similar to the one considered for Schrödinger operators in [5, Theorem 8.3]. In Section 5.2 we show that the necessary conditions for applying the algorithm from Chapter 3 are satisfied. In particular, we construct an orthonormal basis whose linear span is a core for the Dirac operator, so that Lemma 3.25 can be applied. Finally, in Section 5.3 we tie everything together and use Proposition 3.29 to show the desired Σ_1-classification.

5.1 The Dirac operator

A physical account of the Dirac equation and the Dirac operator is beyond the scope of this thesis. We will be simply consider its mathematical aspects, referring to [31] for the physically inclined reader. We will work inside the Hilbert space $\mathcal{H} \coloneqq (L^2(\mathbb{R}^3))^4 \cong L^2(R^3; \mathbb{C}^4)$ whose elements are four-component column vectors $\psi = (\psi_1, \psi_2, \psi_3, \psi_4)^T$ with $\psi_i \in L^2(\mathbb{R}^3)$. The inner product is just defined by summing over the four components,

$$\langle \psi, \varphi \rangle_{\mathcal{H}} = \sum_{i=1}^{4} \langle \psi_i, \varphi_i \rangle_{L^2} = \sum_{i=1}^{4} \int_{\mathbb{R}^3} \psi_i(\mathbf{x}) \overline{\varphi_i(\mathbf{x})} \, \mathrm{d}\mathbf{x} = \int_{\mathbb{R}^3} \sum_{i=1}^{4} \psi_i(\mathbf{x}) \overline{\varphi_i(\mathbf{x})} \, \mathrm{d}\mathbf{x}.$$

The norm on $\mathcal{H}$ is then given by

$$\|\psi\|^2 = \langle \psi, \psi \rangle_{\mathcal{H}} = \sum_{k=1}^{4} \langle \psi_k, \psi_k \rangle_{L^2} = \|\psi_1\|^2 + \|\psi_2\|^2 + \|\psi_3\|^2 + \|\psi_4\|^2.$$

Sometimes we will add a subscript $\mathcal{H}$ or L^2 to norms and inner products for emphasis but the space should be generally clear from the context.

The free Dirac operator

The free Dirac operator H_0 is formally defined in $\mathcal{H}$ by [31]

$$H_0 \psi = -i \, \boldsymbol{\alpha} \cdot \nabla \psi + \beta \psi. \tag{5.1}$$

Here, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \alpha_3)$ where each $\alpha_i, \beta \in \mathbb{C}^{4\times 4}$ are Hermitian matrices which are defined by the *Pauli* matrices,

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

Specifically, they can be written in block matrix form as

$$\alpha_i = \begin{pmatrix} \mathbf{0} & \sigma_i \\ \sigma_i & \mathbf{0} \end{pmatrix}, \quad \beta = \begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & -\mathbf{1} \end{pmatrix},$$

where $\mathbf{1}$ and $\mathbf{0}$ are the identity and zero matrices in $\mathbb{C}^{2\times 2}$, respectively. The gradient operator $\nabla = (\nabla_1, \nabla_2, \nabla_3)$ acts on ψ by $\nabla_i \psi = (\partial_i \psi_1, \partial_i \psi_2, \partial_i \psi_3, \partial_i \psi_4)^T$ where $\partial_i \psi_k$ is shorthand for $\partial \psi_k / \partial x_i$. Thus for $j = 1, 2, 3, 4$, (5.1) can be written component-wise as

$$(H_0 \psi)_j(\mathbf{x}) = -i \sum_{i=1}^{3} \sum_{k=1}^{4} (\alpha_i)_{jk} \frac{\partial \psi_k}{\partial x_i} + \sum_{k=1}^{4} \beta_{jk} \psi_k(\mathbf{x}). \tag{5.2}$$

Being a first order differential operator, we must choose an appropriate dense subspace of $\mathcal{H}$ as the domain of H_0. We clearly need some regularity on the domain for (5.1) to make sense. In [31, pp. 11–12] it is shown that H_0 is essentially self-adjoint on Schwartz space $\mathcal{S}(\mathbb{R}^3)^4$ and that its (self-adjoint) closure has as its domain the first Sobolev space $H^1(\mathbb{R}^3)^4$, so we will take this to be the domain on the free Dirac operator H_0 (for more on the spaces $\mathcal{S}$ and H^1, see e.g. [15]). Note that this means that Schwartz space is a core for H_0. The spectrum of H_0 equals $(-\infty, 1] \cup [1, \infty)$, as shown in [31].

Dirac Operator with potential

The free Dirac operator describes a particle moving freely through empty space, but many applications include for example an electromagnetic field that exerts a force and thus alters the potential energy of the particle. Mathematically this corresponds to the addition of a *potential* function V to H_0:

$$H = H_0 + V. \tag{5.3}$$

In general, the potential is given by a 4×4 matrix-valued function $V(\mathbf{x}) = (V_{ij}(\mathbf{x})), i, j = 1, \ldots, 4$ and acts as a multiplication operator on $L^2(\mathbb{R}^3)^4$. We want a self-adjoint operator, so in particular the matrix $V(\mathbf{x})$ must be Hermitian for all $\mathbf{x} \in \mathbb{R}^3$. For a number of different potentials of physical interest, see for example [31, Chapter 4]. We will take V to mean the multiplication operator acting on $\mathcal{H}$. In this chapter we will assume that each $V_{ij} \in L^2(\mathbb{R}^3)$ is *essentially bounded* with respect to the Lebesgue measure μ in the sense that

$$\|V_{ij}\|_\infty \coloneqq \inf\{M \geq 0 : \mu(\{\mathbf{x} : |V_{ij}(\mathbf{x})| > M\}) = 0\} < \infty.$$

Proposition 5.1. *The multiplication operator V is bounded.*

Proof. First note that V is everywhere defined on $\mathcal{H}$ since for any $\psi \in L^2(\mathbb{R}^3)^4$,

$$\|V\psi\|_{\mathcal{H}}^2 = \sum_{j=1}^{4} \langle (V\psi)_j, (V\psi)_j \rangle \tag{5.4}$$

$$\langle (V\psi)_j, (V\psi)_j \rangle = \int_{\mathbb{R}^3} \left(\sum_{k=1}^{4} V_{jk}\psi_k \right) \overline{\left(\sum_{k=1}^{4} V_{jk}\psi_k \right)} \, \mathrm{d}\mathbf{x}. \tag{5.5}$$

Since the functions V_{ij} are bounded almost everywhere by a constant, each of the four inner products on the form (5.5) in (5.4) is bounded by a linear combination of inner products on the form $\langle \psi_j, \psi_k \rangle_{L^2}$ which are of course finite. Hence $\|V\psi\|_{\mathcal{H}}^2 < \infty$ and so $V\psi \in \mathcal{H}$ for all $\psi \in \mathcal{H}$. Since $V(\mathbf{x})$ is Hermitian for all $\mathbf{x}$, V is symmetric. Thus by the Hellinger-Topelitz theorem, V is a bounded operator on $\mathcal{H}$. ■

Corollary 5.2. *If V is bounded and self-adjoint, and H_0 is self-adjoint on $H^1(\mathbb{R}^3)^4$ it follows from the Kato-Rellich theorem [25, p. 162] that $H = H_0 + V$ is self-adjoint on $H^1(\mathbb{R}^3)^4$.*

Finally, we make a minor additional assumption on V, as done in similar work on Schrödinger operators in [5]. Let $\phi : [0, \infty) \to [0, \infty)$ be an increasing function. Our domain of operators are all Dirac operators $H = H_0 + V$ in the following set:

$$\Omega_\phi = \left\{ H \,:\, \mathcal{D}(H) = H^1(\mathbb{R}^3)^4,\ V_{ij} \in \mathrm{BV}_\phi(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3),\ 1 \le i, j \le 4 \right\} \tag{5.6}$$

where

$$\mathrm{BV}_\phi(\mathbb{R}^3) = \left\{ f : \mathrm{TV}\left(f_{[-a,a]^3}\right) \le \phi(a) \right\}, \tag{5.7}$$

$f_{[-a,a]^3}$ is the restriction of f to the cube $[-a, a]^3$, and TV is the total variation in the sense of Hardy and Krause (see Chapter 4). For $r > 0$ let $M([-r, r]^3)$ be the set of measureable functions on $[-r, r]^3$ and define the space

$$A_r \coloneqq \{ f \in M([-r, r]^3) : \|f\|_\infty + \mathrm{TV}_{[-r,r]^3}(f) < \infty \}. \tag{5.8}$$

This space is a Banach algebra with the norm $\|f\|_{\mathcal{A}_r} = \|f\|_\infty + \sigma \mathrm{TV}_{[-r,r]^3}(f)$ where $\sigma = 3^3 + 1$ [6].

5.2 Approximating the inverse resolvent norm from potential point samples

The goal is to apply Proposition 3.29, in which the key tool is a sequence of computable functions γ_n that converges *uniformly from above* to the reciprocal of the resolvent norm $\gamma(z, A)$ on compact subsets of $\mathbb{C}$. The goal in this section is to construct such a sequence of functions.

Idea behind the construction

By Lemma 3.25, if $\{\psi_n\}_{n=1}^\infty$ is a basis whose linear span forms a core for H, then the functions

$$\Psi_n(z, H) \coloneqq \sigma_1((H - zI)P_n),$$

where P_n is the orthogonal projection onto $\mathrm{Span}\{\psi_1, \ldots, \psi_n\}$, converges uniformly from above $\gamma(z, H) = \sigma_1(H - zI) = \|R(z, H)\|^{-1}$. Thus the question becomes how to approximate $\Psi_n(z, H)$ using arithmetic operations and comparisons. As shown in [8], we have the following basic result which gives the required reduction to computable finite dimensional problems (in the self-adjoint case):

Lemma 5.3. *Let $n \in \mathbb{N}$ and $\varepsilon > 0$ be given. Suppose we can compute matrix $W_n(z) \in \mathbb{C}^{n\times n}$ with entries*

$$W_n(z)_{ij} = \langle (H - zI)\psi_j, (H - zI)\psi_i\rangle + E^n_{ij}$$

for $1 \le i, j \le n$ where the entrywise errors E^n_{ij} have magnitude less than or equal to ε. Then

$$|\Psi_n(z, H)^2 - \sigma_1(W_n)| \le n\varepsilon,$$

and from this it follows that we can compute $\Psi_n(z, H)^2$ to an accuracy of $2n\varepsilon$.

Proof. Given W_n (we drop the reference to z to simplify notation), the matrix with entries

$$\frac{1}{2}\left(\{W_n\}_{ij} + \overline{\{W_n\}_{ji}}\right)$$

is Hermitian and still has entrywise error less than ε. Hence we may assume without loss of generality that W_n is self-adjoint. We denote the matrix without errors by $\widetilde{W}_n$ and view $(H - zI)P_n$ as a linear operator acting on the finite dimensional space $P_n(\mathcal{H}) = \mathrm{Span}\{\psi_1, \ldots, \psi_n\}$. Clearly this matrix is Hermitian, and by definition $\widetilde{W}_n$ is just the matrix representation of the operator $((H-zI)P_n)^*(H-zI)P_n$ acting on the finite dimensional space $P_n(\mathcal{H})$ with respect to the basis $\{\psi_1, \ldots, \psi_n\}$. Thus $\widetilde{W}_n$ is positive semi-definite since

$$\min_{x\in\mathbb{C}^n} \langle \widetilde{W}_n x, x\rangle = \min_{\psi\in P_n(\mathcal{H})} \langle (H - zI)\psi, (H - zI)\psi\rangle \ge 0.$$

By definition, $\sigma_1(\widetilde{W}_n)$ is the square root of the smallest eigenvalue of the positive semi-definite matrix $\widetilde{W}_n^*\widetilde{W}_n = \widetilde{W}_n^2$, i.e., it is simply the smallest eigenvalue of $\widetilde{W}_n$. On the other hand, $\sigma_1((H - zI)P_n)^2$ is the smallest eigenvalue of $((H - zI)P_n)^*(H - zI)P_n$ so we see that

$$\sigma_1(\widetilde{W}_n) = \sigma_1((H - zI)P_n)^2 = \Psi_n(z, H)^2.$$

Using the fact that $|\sigma_1(A) - \sigma_1(B)| \le \|A - B\|$ for matrices $A, B \in \mathbb{C}^{n\times n}$, we have

$$|\sigma_1(W_n) - \Psi_n(z, H)^2| = |\sigma_1(W_n) - \sigma(\widetilde{W}_n)| \le \|W_n - \widetilde{W}_n\| \le n\varepsilon, \qquad (5.9)$$

where the last inequality holds because for a finite matrix M, the operator norm $\|M\|$ is bounded above by its Frobenius norm $\sqrt{\sum |M_{ij}|^2}$. As shown in Proposition A.1, we can compute an approximation $\widehat{\sigma}_1(W_n)$ of $\sigma_1(W_n)$ to any precision using finitely many arithmetic operations and comparisons, including

$$|\widehat{\sigma}_1(W_n) - \sigma_1(W_n)| \le n\varepsilon.$$

Combining with (5.9) we finally have

$$|\widehat{\sigma}_1(W_n) - \Psi_n(z, H)^2| \le 2n\varepsilon,$$

and the lemma follows. ∎

Next, the plan is to construct a basis $\{\psi_n\}_{n=1}^{\infty}$ whose span is a core for H. Then we will show that the inner products

$$\langle (H - zI)\psi_n, (H - zI)\psi_m \rangle \tag{5.10}$$

can be computed from n and m to arbitrary precision using finitely many arithmetic operations and comparisons. Throughout we let H refer to an arbitrary operator in Ω_ϕ.

Remark 5.4. When using the terms *compute* or *approximate* in the following, this refers to a computation done by a finite number of arithmetic operations and comparisons.

Choice of basis

First, consider the Hilbert space $L^2(\mathbb{R})$. In this space, the *Hermite functions* $\{h_n\}_{n=0}^{\infty}$ are defined by

$$h_n(x) = \left(2^n n! \sqrt{\pi}\right)^{-1/2} \exp(-x^2/2) H_n(x), \tag{5.11}$$

where H_n is the n-th Hermite polynomial,

$$H_n(x) = (-1)^n \exp(x^2) \frac{\mathrm{d}^n}{\mathrm{d}x^n} \exp(-x^2).$$

The Hermite functions h_n form an orthonormal basis for $L^2(\mathbb{R})$ [13, p. 187] and satisfy the recurrence relations

$$h_n'(x) = \sqrt{\frac{n}{2}} h_{n-1}(x) - \sqrt{\frac{n+1}{2}} h_{n+1}(x), \tag{5.12}$$

$$x h_n(x) = \sqrt{\frac{n}{2}} h_{n-1}(x) + \sqrt{\frac{n+1}{2}} h_{n+1}(x). \tag{5.13}$$

We will also make use of a classical inequality which says that for the n-th Hermite polynomial H_n we have $H_n(x) \leq (2^n n!)^{1/2} \exp(x^2/2)$ for all $x \in \mathbb{R}$ [17]. Equivalently, for the n-th Hermite function h_n defined in (5.11) we have

$$h_n(x) \leq \pi^{-1/4} < 1. \tag{5.14}$$

Let $\mathbb{Z}_{\geq 0}$ denote the set of all non-negative integers and let $\mathbb{Z}_{\geq 0}^3$ denote the set of all length three multi-indices, $m = (m_1, m_2, m_3), m_i \in \mathbb{Z}_{\geq 0}$. For each multi-index $m \in \mathbb{Z}_{\geq 0}^3$ we will use the notation $|m| = m_1 + m_2 + m_3$. Taking products of Hermite functions, we obtain an orthonormal basis $\{h_m\}_{m \in \mathbb{Z}_{\geq 0}^3}$ for $L^2(\mathbb{R}^3)$,

$$h_m(\mathbf{x}) \coloneqq h_{m_1}(x_1) h_{m_2}(x_2) h_{m_3}(x_3). \tag{5.15}$$

To enumerate these products over $\mathbb{N}$, we need an injective map from $\mathbb{Z}_{\geq 0}^3$ to $\mathbb{N}$. A simple way to obtain this is to consider the sets $S_n \coloneqq \{h_m \in \mathbb{Z}_{\geq 0}^3 \colon |m| \leq n\}$ for all $n \geq 0$. Let r_n be the number of elements in S_n. By a "stars and bars"-argument $r_n = \sum_{k=0}^{n} \binom{k+2}{2} = (n+1)(n+2)(n+3)/6$. The set S_0 contains only the function $h_{0,0,0}(\mathbf{x}) = \pi^{3/2} \exp\left((-x_1^2 - x_2^2 - x_3^2)/2\right)$ which we label e_1.

S_1 contains S_0 and all three functions on the form (5.15) where exactly one of the m_i is 1 and the other two are zero, which we list as $\{e_2, e_3, e_4\}$ (the precise order is immaterial). In general, given an enumeration $\{e_1, \dots, e_{r_n}\}$ of S_n we enumerate the remaining elements of S_{n+1} as $\{e_{r_n+1}, \dots, e_{r_{n+1}}\}$. By this inductive method we obtain an enumeration of all functions on the form (5.15) by $\mathbb{N}$, which we denote by $\mathcal{B} \coloneqq \{e_n\}_{n=1}^{\infty} = \{h_m\}_{m\in\mathbb{Z}^3_{\geq 0}}$.

Of course this may be a computationally suboptimal enumeration for a given application, but it does certainly define a computable (by finite arithmetic means) bijective map from $\mathbb{Z}^3_{\geq 0}$ to $\mathbb{N}$ which we denote by ζ, and its inverse by $\xi : \mathbb{N} \to \mathbb{Z}^3_{\geq 0}$. Hence the list $\{e_n\}_{n=1}^{\infty}$ is an enumeration of the orthonormal basis $\{h_m\}_{m\in\mathbb{Z}^3_{\geq 0}}$ of $L^2(\mathbb{R}^3)$ with $h_m = e_{\zeta(m)}$ for each multi-index $m \in \mathbb{Z}^3_{\geq 0}$ and $e_n = h_{\xi(n)}$ for each $n \in \mathbb{N}$. Note that by construction, $\zeta(m) > |m|$ for all $m \in \mathbb{Z}^3_{\geq 0}$ and so $|\xi(n)| < n$ for any n. This will be important in the proof of Lemma 5.5 below, showing that the basis is a core. To obtain an orthonormal basis for $\mathcal{H}$, let $\{v_1, v_2, v_3, v_4\}$ denote the standard basis in four-dimensional Euclidean space, and for each $n \in \mathbb{N}$ let

$$\psi_n \coloneqq e_{\lceil n/4 \rceil} v_{(n \bmod 4)+1}. \tag{5.16}$$

Let this basis be denoted as $\mathcal{B}_{\mathcal{H}} \coloneqq \{\psi_n\}_{n=1}^{\infty}$. We have

$$\psi_1 = \begin{pmatrix} e_1 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \quad \psi_2 = \begin{pmatrix} 0 \\ e_1 \\ 0 \\ 0 \end{pmatrix}, \quad \psi_3 = \begin{pmatrix} 0 \\ 0 \\ e_1 \\ 0 \end{pmatrix}, \quad \psi_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ e_1 \end{pmatrix},$$

$\psi_5 = e_2 v_1, \dots, \psi_8 = e_2 v_4$ and so forth. Since $\{e_n\}_{n=1}^{\infty}$ is an orthonormal basis for $L^2(\mathbb{R}^3)$, the set $\{\psi_n\}_{n=1}^{\infty}$ is an orthonormal basis for $\mathcal{H}$.

Finally, we need to show that $\mathcal{S} \coloneqq \operatorname{Span}\{\psi_n : n \in \mathbb{N}\}$ is a core for H. Since $H = -i\boldsymbol{\alpha} \cdot \nabla + \beta + V$ and $\beta + V$ is a bounded, hence continuous operator on all of $\mathcal{H}$, every dense subset is a core for $\beta + V$. Thus we can neglect those terms when showing that $\mathcal{S}$ is a core for H. For the proof we use the same idea as in [8, Proposition 7.1].

Lemma 5.5. *Let $H = -i\boldsymbol{\alpha} \cdot \nabla + \beta + V$ be the self-adjoint Dirac operator with domain $H^1(\mathbb{R}^3)^4$. Then $\mathcal{S} = \operatorname{Span}\{\psi_n : n \in \mathbb{N}\}$ is a core for H.*

Proof. As explained, we omit the term $\beta + V$ from the formal definition of H. We know from [31, Theorem 1.1] that $-i\boldsymbol{\alpha} \cdot \nabla$ defined on $\mathcal{S}(\mathbb{R}^3)^4$ is essentially self-adjoint and that its unique self-adjoint extension has domain $H^1(\mathbb{R}^3)^4 \supset \mathcal{S}$. Let T denote the closure of the formal operator $-i\boldsymbol{\alpha} \cdot \nabla$ acting on the domain $\mathcal{S}$. If we can show that $\mathcal{S}(\mathbb{R}^3)^4 \subset \mathcal{D}(T)$, then $H \restriction \mathcal{S}(\mathbb{R}^3)^4 \subset T \subset H$ and taking closures it follows that $T = H$. Thus $\mathcal{S}$ is a core for $-i\boldsymbol{\alpha} \cdot \nabla$ defined on $H^1(\mathbb{R}^3)^4$, and hence it is also a core for H on the same domain.

Let $g \in \mathcal{S}(\mathbb{R}^3)^4$. Since $\{\psi_n\}_n$ is an orthonormal basis for $\mathcal{H}$ we may write

$$g = \sum_{n=1}^{\infty} \langle g, \psi_n \rangle \psi_n.$$

Defining $g_m \coloneqq \sum_{n=1}^{m} \langle g, \psi_n \rangle \psi_n$ we have $\mathcal{D}(T) \supset \mathcal{S} \ni g_m \to g$. Because T is closed, if Tg_m converges it follows that $g \in \mathcal{D}(T)$. Now

$$Tg_m = \sum_{n=1}^{m} \langle g, \psi_n \rangle T\psi_n = -i \sum_{n=1}^{m} \langle g, \psi_n \rangle (\boldsymbol{\alpha} \cdot \nabla)\psi_n,$$

so we see that Tg_m converges if and only if the limit

$$\lim_{m \to \infty} \sum_{n=1}^{m} \langle g, \psi_n \rangle (\boldsymbol{\alpha} \cdot \nabla)\psi_n$$

exists in $\mathcal{H}$. Since absolute convergence implies convergence in Hilbert space, the above limit exists if the positive series

$$\sum_{n=1}^{\infty} |\langle g, \psi_n \rangle| \, \|(\boldsymbol{\alpha} \cdot \nabla)\psi_n\| \tag{5.17}$$

converges in $\mathbb{R}$. We will prove this in two parts by showing that:

1. The sequence $|\langle g, \psi_n \rangle|$ tends to zero faster than the reciprocal of any polynomial in n.
2. The norms $\|(\boldsymbol{\alpha} \cdot \nabla)\psi_n\|$ grow only polynomially in n.

For the first part of the claim, note that $g = (g_1, g_2, g_3, g_4)^T$ where each $g_i \in \mathcal{S}(\mathbb{R}^3)$. By definition of ψ_n, for each $n \in \mathbb{N}$

$$\langle g, \psi_n \rangle_{\mathcal{H}} = \left\langle g_{(n \bmod 4)+1}, e_{\lceil n/4 \rceil} \right\rangle_{L^2(\mathbb{R}^3)}.$$

So the sequence $\langle g, \psi_n \rangle_{\mathcal{H}}$ is just a sequence of inner products in $L^2(\mathbb{R}^3)$,

$$\langle g_1, e_1 \rangle, \langle g_2, e_1 \rangle, \langle g_3, e_1 \rangle, \langle g_4, e_1 \rangle, \langle g_1, e_2 \rangle, \langle g_2, e_2 \rangle, \ldots$$

The sequence of inner products $\langle g, \psi_n \rangle_{\mathcal{H}}$ consists of exactly four subsequences of inner products in $L^2(\mathbb{R}^3)$, one for each coordinate of g, with a fixed distance of exactly four terms between consecutive elements from each subsequence. If we can show that each of the four subsequences decay faster than any polynomial in n, then so will the entire sequence as well. So let us consider the subsequence corresponding to the first coordinate of g, i.e., the sequence of inner products $\langle g_1, e_n \rangle, n \in \mathbb{N}$. Recall the bijection $\xi : \mathbb{N} \to \mathbb{Z}^3_{\geq 0}$ from the enumeration we picked for the e_n so that $e_n = h_{\xi(n)}$ for each $n \in \mathbb{N}$. By construction, $|\xi(n)|$ grows polynomially in n. Since $g_1 \in \mathcal{S}(\mathbb{R}^3) \subset L^2(\mathbb{R}^3)$ we can write

$$g_1 = \sum_{n=1}^{\infty} \langle g_1, e_n \rangle e_n. \tag{5.18}$$

Next we will employ a trick using the Harmonic oscillator in $L^2(\mathbb{R}^3)$ formally defined by

$$D = \sum_{i=1}^{3} \left(-\frac{\partial^2}{\partial x_i^2} + x_i^2 \right).$$

It is well known that D is essentially self-adjoint on $\mathcal{S}(\mathbb{R}^3)$ and that for $m \in \mathbb{Z}^3_{\geq 0}$, $h_m \in L^2(\mathbb{R}^3)$ is an eigenvector with $Dh_m = (2|m| + 3)h_m$ (see Example 2.13 and the corresponding reference). Thus $De_n = (2|\xi(n)| + 3)e_n$ for each $n \in \mathbb{N}$. Also note that $g_1 \in \mathcal{D}(D^k)$ for any $k \in \mathbb{N}$ since Schwartz space is closed under differentiation and multiplication by polynomials. Then we have

$$\langle Dg_1, e_n \rangle = \langle g_1, De_n \rangle = (2|\xi(n)| + 1)\langle g_1, e_n \rangle$$

and since $Dg_1 \in L^2(\mathbb{R}^3)$, it follows that the sequence $\{(2|\xi(n)| + 3)|\langle g_1, e_n \rangle|\}_{n=1}^{\infty}$ is square summable in $\mathbb{R}$. Since $g_1, e_n \in \mathcal{D}(D^k)$ for all k, we can repeat this argument an arbitrary number of times with $D^2, D^3, \ldots$, to see that the sequence $\{|\langle g_1, e_n \rangle|\}_{n=1}^{\infty}$ must tend towards zero with n faster than the reciprocal of any polynomial in n. Using the same exact argument for the three other subsequences $\langle g_i, e_n \rangle$ for $i = 2, 3, 4$, we have shown the first part of the claim.

For the second part of the claim, arguing as above by considering one subsequence for each of the four components of the wavefunction, it is sufficient to prove that the sequence of norms $\|(\boldsymbol{\alpha} \cdot \nabla)e_n v_1\|$ grows polynomially in n. Observe that for any $h_m(\mathbf{x}) = h_{m_1}(x_1)h_{m_2}(x_2)h_{m_3}(x_3)$, using orthonormality, the recurrence relation (5.12) and the bound (5.14), we have

$$\begin{aligned}
\|\partial_i h_m\|_{L^2(\mathbb{R}^3)} &= \left\|h'_{m_i}\right\|_{L^2(\mathbb{R})} \\
&= \left\| \sqrt{\frac{m_i}{2}} h_{m_i - 1} - \sqrt{\frac{m_i + 1}{2}} h_{m_i + 1} \right\|_{L^2(\mathbb{R})} \\
&< \sqrt{2(m_i + 1)} \leq \sqrt{2|m| + 1}.
\end{aligned}$$

Recall that the columns of each matrix $\alpha_i \in \mathbb{C}^{4\times 4}$ have exactly one entry of modulus one, and the other three equal zero. Thus by the above inequality, for $i = 1, 2, 3$ and $n \in \mathbb{N}$

$$\|\alpha_i \nabla_i e_n v_1\|_{\mathcal{H}} = \|\alpha_i \partial_i e_n v_1\|_{\mathcal{H}} = \|\partial_i h_{\xi(n)}\|_{L^2(\mathbb{R}^3)} < \sqrt{2|\xi(n)| + 1},$$

and so

$$\|(\boldsymbol{\alpha} \cdot \nabla)e_n v_1\|_{\mathcal{H}} = \|(\alpha_1 \nabla_1 + \alpha_2 \nabla_2 + \alpha_3 \nabla_3)e_n v_1\|_{\mathcal{H}} < 3\sqrt{2|\xi(n)| + 1}.$$

From this the second part of the claim follows by using the same exact argument for the sequences $\|(\boldsymbol{\alpha} \cdot \nabla)e_n v_i\|_{\mathcal{H}}$ with $i = 2, 3, 4$. Thus we have shown that the series (5.17) converges in $\mathbb{R}$ and so the sequence Tg_m converges in $\mathcal{H}$, finishing the proof. ■

Computing inner products

Having shown that our chosen basis is a core, the next step is to show that we can actually approximate the matrices in Lemma 5.3 to an arbitrary level of precision in each entry. This will follow if we can show that we can compute the inner products (5.10) to an arbitrary level of precision. We first expand the inner products, using the self-adjointness of H:

$$\begin{aligned}
&\langle (H - zI)\psi_n, (H - zI)\psi_m \rangle \\
&= \langle H\psi_n, H\psi_m \rangle - z\langle H\psi_n, \psi_m \rangle - \bar{z}\langle H\psi_n, \psi_n \rangle + z\bar{z}\langle \psi_n, \psi_m \rangle
\end{aligned}$$

$$= \langle H\psi_n, H\psi_m\rangle - 2\,\mathrm{Re}(z)\langle H\psi_n, \psi_m\rangle + |z|^2\langle \psi_n, \psi_m\rangle.$$

The inner products $\langle \psi_n, \psi_m\rangle_{\mathcal{H}}$ are trivial to compute so we only have to consider the inner products

$$\langle H\psi_n, H\psi_m\rangle = \langle H_0\psi_n, H_0\psi_m\rangle + \langle H_0\psi_n, V\psi_m\rangle + \langle V\psi_n, H_0\psi_m\rangle + \langle V\psi_n, V\psi_m\rangle,$$
$$\langle H\psi_n, \psi_m\rangle = \langle H_0\psi_n, \psi_m\rangle + \langle V\psi_n, \psi_m\rangle.$$

If we can show that the above inner products can be computed to arbitrary precision, then we know that the conditions in Lemma 5.3 are satisfied. First we deal with the inner products only containing H_0.

Proposition 5.6. *Let H_0 be the free Dirac operator acting on $H^1(\mathbb{R}^3)^4$. Then for any $\psi_n, \psi_m \in \mathcal{B}_{\mathcal{H}}$, the inner products $\langle H_0\psi_n, H_0\psi_m\rangle$ and $\langle H_0\psi_n, \psi_m\rangle$ can be computed to arbitrary precision.*

Proof. Let us first focus on how H_0 acts on the basis functions ψ_n. Suppose $\psi_n = h_m v_1$ for some $m \in \mathbb{Z}^3_{\geq 0}$, i.e.,

$$\psi(\mathbf{x}) = \begin{pmatrix} h_m(\mathbf{x}) \\ 0 \\ 0 \\ 0 \end{pmatrix},$$

where $h_m(\mathbf{x}) = h_{m_1}(x_1)h_{m_2}(x_2)h_{m_3}(x_3)$. Recall from (5.2) that H_0 works component-wise on $\psi_n = h_m v_1$ which has (writing component index in superscript) $\psi_n^1 = h_m$ and $\psi_n^k = 0$ for $k = 2, 3, 4$, by

$$\begin{aligned}(H_0\psi_n)_j(\mathbf{x}) &= -i\sum_{i=1}^{3}\sum_{k=1}^{4}(\alpha_i)_{jk}\frac{\partial\psi_n^k}{\partial x_i}(\mathbf{x}) + \sum_{k=1}^{4}\beta_{jk}\psi_n^k(\mathbf{x}) \\ &= -i\sum_{i=1}^{3}(\alpha_i)_{j1}\frac{\partial\psi_n^1}{\partial x_i}(\mathbf{x}) + \beta_{j1}\psi_n^k(\mathbf{x}) \\ &= -i\left((\alpha_1)_{j1}\frac{\partial h_m}{\partial x_1}(\mathbf{x}) + (\alpha_2)_{j1}\frac{\partial h_m}{\partial x_2}(\mathbf{x}) + (\alpha_3)_{j1}\frac{\partial h_m}{\partial x_3}(\mathbf{x})\right) + \beta_{j1}h_m(x).\end{aligned}$$

Since $h_m(\mathbf{x}) = h_{m_1}(x_1)h_{m_2}(x_2)h_{m_3}(x_3)$ we have

$$\frac{\partial h_m}{\partial x_1} = h'_{m_1}(x_1)h_{m_2}(x_2)h_{m_3}(x_3). \tag{5.19}$$

By the recurrence relation (5.12) for the Hermite functions in $L^2(\mathbb{R})$,

$$h'_{m_1}(x_1) = \sqrt{\frac{m_1}{2}}h_{m_1-1}(x_1) - \sqrt{\frac{m_1+1}{2}}h_{m_1+1}(x_1), \tag{5.20}$$

Using this in (5.19), we see that $\partial e_m/\partial x_1$ is a linear combination of the basic tensor product of Hermite functions in $L^2(\mathbb{R}^3)$. Similarly, $\partial e_m/\partial x_2$ and $\partial e_m/\partial x_3$ are also just linear combinations of the functions $\{e_m\}$. The coefficients in these linear combinations can be computed in finite time via the recurrence relations, assuming square roots of integers can be computed to arbitrary precision. Finally, using these observations in the expression obtained

for $(H_0\psi_n)_j(\mathbf{x})$ above, this shows that each component function $(H_0\psi_n)_j$ of $H_0\psi_n$ is a finite linear combination of functions in $\mathcal{B}$, where the coefficients can be calculated from n.

Clearly the above arguments also hold if $\psi_n = e_m v_i$ for $i \in \{2,3,4\}$ as well, and so it holds for all the basis functions in $\mathcal{B}_\mathcal{H}$. Thus by the definition of the inner product in $\mathcal{H}$ as summing the inner product over the four components, the inner product $\langle H_0\psi_n, H_0\psi_m\rangle$ is a linear combination of inner products in $L^2(\mathbb{R}^3)$ on the form $\langle e_n, e_m\rangle = \delta_{nm}$ with $e_n, e_m \in \mathcal{B}$. In particular, so are the inner products $\langle H_0\psi_n, \psi_m\rangle$. Since the coefficients in the linear combinations can be calculated from n and m by the Hermite recurrence relations and the constants $(\alpha_\ell)_{jk}, \beta_{jk}$ are known, the lemma follows. ■

Since we showed above that for any $\psi_n \in \mathcal{B}_\mathcal{H}$, $H_0\psi_n$ can always be written as a finite linear combination of other basis functions in $\mathcal{B}_\mathcal{H}$, by the linearity of the inner product, we can without loss of generality replace each occurrence of $H_0\psi_n$ with simply ψ_n. Since V is self-adjoint, it remains to show that the inner products $\langle V\psi_n, V\psi_m\rangle$ and $\langle V\psi_n, \psi_m\rangle$ can be computed with arbitrary precision by finite arithmetic means. Let us first consider how V acts on ψ_n. As above, we assume that $\psi_n = e_m v_1$ for some $m \in \mathbb{Z}^3_{\geq 0}$. In this case, the four components of $V\psi_n$ are given by

$$(V\psi_n)_j(\mathbf{x}) = \sum_{k=1}^{4} V_{jk}(\mathbf{x})\psi_n^k(\mathbf{x}) = V_{j1}(\mathbf{x})\psi_n^1(\mathbf{x}) = V_{j1}(\mathbf{x})e_m(\mathbf{x}),$$

i.e., each component of $V\psi_n$ is a product of a function V_{ij} with e_m. This is also clearly the case if $\psi_n = e_m v_i$ with $i \in \{2,3,4\}$, which covers all the basis elements in $\mathcal{B}_\mathcal{H}$. Thus the inner product

$$\langle V\psi_n, V\psi_m\rangle_\mathcal{H} = \sum_{j=1}^{4} \int_{\mathbb{R}^3} (V\psi_n)_j(\mathbf{x})\overline{(V\psi_m)_j(\mathbf{x})}\,\mathrm{d}\mathbf{x} \tag{5.21}$$

is a sum of four inner products in $L^2(\mathbb{R}^3)$ on the form

$$\langle f\,e_n, g\,e_m\rangle_{L^2(\mathbb{R}^3)} = \int_{\mathbb{R}^3} f(\mathbf{x})\overline{g(\mathbf{x})}e_n(x)e_m(\mathbf{x})\,\mathrm{d}\mathbf{x} \tag{5.22}$$

where $f, g \in \mathrm{BV}_\phi(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3), n, m \in \mathbb{N}$ and the coefficients can be computed to arbitrary precision from the recurrence relations. If we can compute inner products on the form (5.22) to arbitrary precision, then we will be done, since this also covers the inner products in $\langle V\psi_n, \psi_m\rangle_\mathcal{H}$ using $g(\mathbf{x}) = 1$ in (5.22). Since we have a one-to-one correspondence $e_{\zeta(m)} \leftrightarrow h_m$ for all $m \in \mathbb{Z}^3_{\geq 0}$ via the bijection ζ, this is equivalent to showing that the computation of inner products on the form $\langle fh_n, gh_m\rangle$ where $n, m \in \mathbb{Z}^3_{\geq 0}$ and $f, g \in \mathrm{BV}_\phi(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$ can be done through finite arithmetic means.

To achieve this, the idea is to consider for $r > 0$, the decomposition of $\mathbb{R}^3$ into the regions $\{\mathbf{x} \in \mathbb{R}^3 \colon |x_i| \geq r \text{ for } i = 1,2,3\}$ and $\{\mathbf{x} \in \mathbb{R}^3 \colon |x_i| < r \text{ for } i = 1,2,3\}$:

$$\langle fh_n, gh_m\rangle = \int_{|x_i|<r} f\,\bar{g}\,h_m\,h_n\,\mathrm{d}\mathbf{x} + \int_{|x_i|\geq r} f\,\bar{g}\,h_m\,h_n\,\mathrm{d}\mathbf{x}, \tag{5.23}$$

and compute an r so large that the second integral vanishes and then use Quasi Monte Carlo integration to compute the first integral with a given accuracy. First, we need an bound in the one-dimensional case.

Lemma 5.7. *Let $h_n \in L^2(\mathbb{R})$ be one of the one-dimensional Hermite functions as defined in (5.11). Given $\delta > 0$ one can compute in finitely many arithmetic operations and comparisons a number $r > 0$ such that*

$$\int_{|x|\geq r} h_n^2 \,\mathrm{d}x < \delta. \tag{5.24}$$

Proof. Recall the recurrence relation (5.13). For any $r \geq 1$ this gives,

$$\begin{aligned}\int_{|x|\geq r} h_n(x)^2 \,\mathrm{d}x &\leq \frac{1}{r^2}\int_{\mathbb{R}} x^2 h_n(x)^2 \,\mathrm{d}x \\ &= \frac{1}{r^2}\int_{\mathbb{R}} \left(\sqrt{\frac{n}{2}}h_{n-1}(x) + \sqrt{\frac{n+1}{2}}h_{n+1}(x)\right)^2 \mathrm{d}x \\ &= \frac{n+1/2}{r^2},\end{aligned}$$

where the last equality follows from orthonormality of the $\{h_n\}_n$. Given $n \in \mathbb{N}$ and $\delta > 0$, it is straightforward to compute a number $r > 0$ (say, the smallest positive integer) such that $(n+1/2)\delta^{-1} < r^2$ and (5.24) holds. ■

We can use this to establish conditions for computing an $r > 0$ such that the second integral in (5.23) becomes arbitrarily small.

Proposition 5.8. *Let f, g, h_m, h_n be given as in (5.23). Then given any $\delta > 0$ one can compute in finite time a number $r > 0$ such that*

$$\int_{|x_i|\geq r} f\,\bar{g}\,h_m\,h_n \,\mathrm{d}\mathbf{x} < \delta. \tag{5.25}$$

Proof. For $r > 0$, $h_n = h_{n_1}h_{n_2}h_{n_3}$ and $h_m = h_{m_1}h_{m_2}h_{m_3}$ we have

$$\int_{|x_i|\geq r} f\,\bar{g}\,h_m\,h_n \,\mathrm{d}\mathbf{x} \leq \|f\|_\infty\|g\|_\infty \int_{|x_i|\geq r} |h_n\,h_m| \,\mathrm{d}\mathbf{x}. \tag{5.26}$$

By Hölder's inequality,

$$\int_{|x_i|\geq r} |h_n\,h_m| \,\mathrm{d}\mathbf{x} \leq \left(\int_{|x_i|\geq r} h_n^2 \,\mathrm{d}\mathbf{x}\right)^{1/2}\left(\int_{|x_i|\geq r} h_m^2 \,\mathrm{d}\mathbf{x}\right)^{1/2}. \tag{5.27}$$

Now since $h_n(\mathbf{x}) = h_{n_1}(x_1)h_{n_2}(x_2)h_{n_3}(x_3)$ we have

$$\int_{|x_i|\geq r} h_n^2 \,\mathrm{d}\mathbf{x} = \int_{|x_1|\geq r} h_{n_1}^2 \,\mathrm{d}x_1 \int_{|x_2|\geq r} h_{n_2}^2 \,\mathrm{d}x_2 \int_{|x_3|\geq r} h_{n_3}^2 \,\mathrm{d}x_3.$$

By Lemma 5.7, for $i = 1, 2, 3$ we can compute via finite arithmetic means numbers $r_i > 0$ such that

$$\int_{|x_i|\geq r_i} h_{n_i}^2 \,\mathrm{d}x_i < \left(\frac{\delta}{\|f\|_\infty\|g\|_\infty}\right)^{1/3}.$$

Taking $r_n = \max\{r_1, r_2, r_3\}$ we then have

$$\int_{|x_i|>r_n} h_n^2 \, \mathrm{d}\mathbf{x} < \frac{\delta}{\|f\|_\infty \|g\|_\infty}.$$

Using the same argument for h_m to compute a corresponding r_m and setting $r = \max\{r_n, r_m\}$, by (5.27) we then have

$$\int_{|x_i|\geq r} |h_n \, h_m| \, \mathrm{d}\mathbf{x} < \frac{\delta}{\|f\|_\infty \|g\|_\infty},$$

and by (5.26), the inequality (5.25) follows. ■

To deal with the first integral in (5.23) we will use the results from Chapter 4 regarding Quasi Monte Carlo Integration. First, we need to establish a bound on the Hardy-Krause variation of the chosen basis functions in $L^2(\mathbb{R}^3)$.

Lemma 5.9. *Let $e_{g(m)}(\mathbf{x}) = h_m(\mathbf{x}) = h_{m_1}(x_1)h_{m_2}(x_2)h_{m_3}(x_3)$ be a basis function in $\mathcal{B}$ and let $r > 0$. Then*

$$\mathrm{TV}_{[-r,r]^3}(h_m) < \left(1 + 2r\sqrt{2(|m|+1)}\right)^3.$$

Proof. First observe that for a one-dimensional Hermite function h_n, the recurrence relation (5.12) and the inequality (5.14) give

$$|h_n'(x)| = \left|\sqrt{\frac{n}{2}}h_{n-1}(x) - \sqrt{\frac{n+1}{2}}h_{n+1}(x)\right| \leq \sqrt{\frac{n}{2}} + \sqrt{\frac{n+1}{2}} < \sqrt{2(n+1)}.$$

Thus for any multi-index I of length k, defining $\tilde{x}$ by $\tilde{x}_i = x_i$ for $i \in I$ and $\tilde{x}_i = r$ for $i \notin I$, another application of (5.14) implies that

$$\begin{aligned}\left|\frac{\partial^k h_m}{\partial x_{i_1} \cdots \partial x_{i_k}}(\tilde{x})\right| &\leq |h'_{m_{i_1}}(x_1)| \cdots |h'_{m_{i_k}}(x_k)| \\ &\leq \sqrt{2(m_{i_1}+1)} \cdots \sqrt{2(m_{i_k}+1)} \\ &\leq \left(\sqrt{2(|m|+1)}\right)^k,\end{aligned}$$

which gives

$$\int_{[-r,r]^k} \left|\frac{\partial^k h_m}{\partial x_{i_1} \cdots \partial x_{i_k}}(\tilde{x})\right| \mathrm{d}x_{i_1} \cdots \mathrm{d}x_{i_k} < \left(2r\sqrt{2(|m|+1)}\right)^k.$$

We are dealing with dimension $d = 3$ and $a = -r$, $b = r$ in (4.4) and so we only need to consider the multi-index sets,

$$\mathcal{I}_1 = \{(1),(2),(3)\}, \quad \mathcal{I}_2 = \{(1,2),(1,3),(2,3)\}, \quad \mathcal{I}_3 = \{(1,2,3)\}.$$

Writing a^k for $\left(2r\sqrt{2(|m|+1)}\right)^k$, the Hardy-Krause variation becomes

$$\mathrm{TV}_{[-r,r]^3}(h_m) = \sum_{k=1}^{3} \sum_{I \in \mathcal{I}_k} \int_{[-r,r]^k} \left|\frac{\partial^k h_m}{\partial x_{i_1} \cdots \partial x_{i_k}}(\tilde{x})\right| \mathrm{d}x_{i_1} \cdots \mathrm{d}x_{i_k}$$

$$\begin{aligned}
&\leq \sum_{k=1}^{3} \sum_{I \in \mathcal{I}_k} \left(2r\sqrt{2(|m|+1)}\right)^k \\
&= 3a + 3a^2 + a^3 \\
&= \left(1 + 2r\sqrt{2(|m|+1)}\right)^3 - 1.
\end{aligned}$$

■

For notational brevity, given $m \in \mathbb{Z}^3_{\geq 0}$ and $r > 0$ we define the quantity

$$L_r(m) \coloneqq \left[1 + \sigma \left(1 + 2r\sqrt{2(|m|+1)}\right)^3\right]$$

where in our case $\sigma = 3^3 + 1$. In the Banach algebra $\mathcal{A}_r$ this means $\|h_m\|_{\mathcal{A}_r} \leq L_r(m)$. Now we are ready to show how to compute the first integral in (5.23):

Proposition 5.10. *Let f, g, h_m, h_n be as in (5.23). Let $\delta > 0$ and let $r > 0$. Then the integral*

$$\int_{|x_i|<r} f\, \bar{g}\, h_m\, h_n \, \mathrm{d}\mathbf{x}$$

can be computed by finite arithmetic means with error less than δ.

Proof. We may assume that $r \in \mathbb{N}$ by rounding upwards if necessary. Since $\mathcal{A}_r$ is a Banach algebra we have

$$\mathrm{TV}_{[-r,r]^3}(f\, \bar{g}\, h_m\, h_n) \leq \|f\, \bar{g}\, h_m\, h_n\|_{\mathcal{A}_r} \leq \phi(r)^2 \cdot L_r(m) \cdot L_r(n),$$

where $\phi(r)$ bounds $\|f\|_{\mathcal{A}_r}$ and $\|g\|_{\mathcal{A}_r}$ as in (5.7). Now compute M large so that

$$(2r)^3 \cdot \frac{C(3)(\log(M)+1)^3}{M} \cdot \phi(r)^2 \cdot L_r(m) \cdot L_r(n) < \delta,$$

where $C(3)$ is computed as in (4.2). Assuming logarithms and square roots can be computed, such an M can be found by finite arithmetic means. Then if $s_k = 2rt_k - (r, r \ldots, r)^T$ are the rescaled Halton points, by Theorem 4.3

$$\left| \frac{(2r)^3}{M} \sum_{k=1}^{M} f(s_k)\, \overline{g(s_k)}\, h_m(s_k)\, h_n(s_k) - \int_{|x_i|<r} f\, \bar{g}\, h_m\, h_n \, \mathrm{d}\mathbf{x} \right| < \delta.$$

Thus we only need to argue that each of the numbers $f(s_k)\, \overline{g(s_k)}\, h_m(s_k)\, h_n(s_k)$ are computable by finite arithmetic means. Since each $s_k \in \mathbb{Q}^3$, by assumption we can compute $f(s_k)\, \overline{g(s_k)}$ exactly. The one-dimensional Hermite functions $h_k(x)$ from (5.16) can be computed to arbitrary precision on rational input $x \in \mathbb{Q}$ by using the recurrence formula (5.12) and e.g. a power series expansion with a Lagrange error bound on the tail for $\exp(-x^2/2)$. By (5.14), $h_m(s_k)h_n(s_k)$ is a product of six numbers each bounded by 1 and computable to arbitrary precision individually. Thus by ensuring the tail error bounds are small enough, $h_m(s_k)h_n(s_k)$is also computable to arbitrary precision. Hence we see that the product $f(s_k)\, \overline{g(s_k)}\, h_m(s_k)\, h_n(s_k)$ can indeed be computed to any level of precision via finitely many arithmetic operations and comparisons. ■

Thus complementing Proposition 5.6 we have:

Proposition 5.11. *Let V be a potential as defined in the setup* (5.6) *acting on $H^1(\mathbb{R}^3)^4$. Then for any $\psi_n, \psi_m \in \mathcal{B}_{\mathcal{H}}$, the inner products $\langle V\psi_n, V\psi_m\rangle$ and $\langle V\psi_n, \psi_m\rangle$ can be computed to arbitrary precision.*

Proof. Recall from (5.21) and (5.22) that $\langle V\psi_n, V\psi_m\rangle$ is a sum of four inner products in $L^2(\mathbb{R}^3)$ on the form $\langle f\, e_n, g\, e_m\rangle$ where $f, g \in \mathrm{BV}_\phi(\mathbb{R}^3) \cap L^\infty(\mathbb{R}^3)$. Let $\delta > 0$ and use Proposition 5.8 to compute an $r \in \mathbb{N}$ such that

$$\int_{|x_i|\geq r} f\,\bar{g}\,h_m\,h_n\,\mathrm{d}\mathbf{x} < \delta/2.$$

Next, Proposition 5.10 says that we can compute

$$\int_{|x_i|< r} f\,\bar{g}\,h_m\,h_n\,\mathrm{d}\mathbf{x}$$

also with error less than $\delta/2$ by finite arithmetic means. Then the decomposition

$$\langle fh_n, gh_m\rangle = \int_{|x_i|< r} f\,\bar{g}\,h_m\,h_n\,\mathrm{d}\mathbf{x} + \int_{|x_i|\geq r} f\,\bar{g}\,h_m\,h_n\,\mathrm{d}\mathbf{x}$$

shows that each inner product $\langle f\, e_n, g\, e_m\rangle$ above can be computed with total error less than δ. Since $\delta > 0$ was arbitrary, it follows that the inner product $\langle V\psi_n, V\psi_m\rangle$, which is a sum of four such inner products, can also be computed to any desired level of precision. This also includes the inner products on the form $\langle V\psi_n, \psi_m\rangle$ since we can use $g(\mathbf{x}) = 1$ in Proposition 5.8 and Proposition 5.10. ∎

Finally, Proposition 5.6 and Proposition 5.11 together prove

Proposition 5.12. *Let $H \in \Omega_\phi$ and $\psi_m, \psi_n \in \mathcal{B}_{\mathcal{H}}$. Then the inner product*

$$\langle (H - zI)\psi_n, (H - zI)\psi_m\rangle$$

can be computed to arbitrary precision with a finite number of arithmetic operations and comparisons.

In conclusion, this means that we can compute to any precision an approximation W_n of the matrix $\widetilde{W}_n$ in Lemma 5.3. Then as shown in Appendix A.1, we can compute $\sigma_1(W_n)$ to any precision, which in turn means that we can estimate $\sigma_1(\widetilde{W}_n) = \Psi_n(z, H)^2$ with arbitrarily small error.

5.3 Main result on Dirac operators with bounded potentials

Having established all the necessary conditions in Lemma 5.3 we can finally tie everything together and show a Σ_1-classification. Since every $H \in \Omega_\phi$ is self-adjoint, in the proof below we can apply Proposition 3.28 and Proposition 3.29, both with $g(x) = x$ in (3.14) which becomes an equality.

Theorem 5.13. *Define the computational problem $\{\Xi, \Omega, \Lambda, \mathcal{M}\}$ as follows.*

(i) *The domain is* Ω_ϕ *as defined in* (5.6).

(ii) *The evaluation functions* Λ *contains the matrix evaluation functions* $f_{i,j}\colon H \mapsto \langle H\psi_j, \psi_i\rangle, g_{i,j}\colon H \mapsto \langle H\psi_j, H\psi_i\rangle$ *with* $\psi_i, \psi_j \in \mathcal{B}_\mathcal{H}$*, the point sampling functions* $f_x(H_0+V) = V(x), x \in \mathbb{Q}^3$ *and the constant functions* $\phi_n\colon H \mapsto \phi(n)$.

(iii) *The problem function is* $\Xi(H) = \mathrm{Sp}(H)$.

(iv) $\mathcal{M}$ *is the collection of non-empty closed subsets of* $\mathbb{C}$ *with the Attouch-Wets metric.*

Then $\{\Xi, \Omega, \Lambda, \mathcal{M}\} \in \Sigma_1$ *as defined in Definition 3.18.*

Proof. Let $H = H_0+V \in \Omega_\phi$ be arbitrary. H is self-adjoint on $\mathcal{D}(H) = H^1(\mathbb{R}^3)^4$ and $\mathrm{Span}\{\psi_1, \psi_2, \ldots\}$ is a core for H by Lemma 5.5. Hence by Lemma 3.25, on compact subsets of $\mathbb{C}$, the sequence of functions

$$\Psi_n(z, H) = \sigma_1((H - zI)P_n)$$

converges uniformly from above to

$$\gamma(z, H) \coloneqq \sigma_1(H - zI) = \|R(z, H)\|^{-1} = \mathrm{dist}(z, \mathrm{Sp}(H)).$$

By Proposition 5.12, we can compute the inner products

$$\langle (H - zI)\psi_m, (H - zI)\psi_n\rangle$$

up to arbitrary precision with finite arithmetic operations and comparisons. Thus given $z \in \mathbb{C}$, for each n we can apply Lemma 5.3 to compute an approximation $v_n(z, H)$ via finite arithmetic means such that

$$\left|v_n(z, H)^2 - \Psi_n(z, H)^2\right| \leq \frac{1}{n^2}. \tag{5.28}$$

Since $\Psi_n(z, H)$ converges locally uniformly to $\gamma(z, H)$, so will $v_n(z, H)$, and the convergence will eventually be monotone. To get convergence from above let

$$\gamma_n(z, H) \coloneqq v_n(z, H) + 1/n.$$

Then γ_n will converge to γ locally uniformly and it follows from (5.28) that $\gamma_n(z, H) \geq \Psi_n(z, H) = \|R(z, H)\|^{-1}$, i.e., the convergence is from above. Hence the functions $\gamma_n(z, H)$ satisfy the assumptions of Proposition 3.29 and so the sets $\Gamma_n(H)$ defined by Algorithm 1 form a height one arithmetic tower which converges to $\mathrm{Sp}(H)$ in the Attouch-Wets topology with Σ_1-error control as described by Definition 3.18. ■

CHAPTER 6

Dirac-Coulomb Operator with Infinite Mass Boundary Conditions

In this chapter we consider a two-dimensional Dirac operator on an infinite sector of the Euclidean plane with an unbounded potential of Coulomb type. Here, the free Dirac operator gives the Hamiltonian of a spin $\frac{1}{2}$-particle confined in a planar region in relativistic motion [19]. Such a two-dimensional model is of interest in the physical study to the study of e.g. electrons in graphene, which is a two-dimensional hexagonal lattice of carbon atoms. For more background, we defer to the introduction in [7] and the references therein.

The goal is to give a Σ_1-classification, following the same strategy as for the three-dimensional Dirac operator with bounded potential in Chapter 5. In Section 6.1 we introduce the two-dimensional Dirac operator and formulate the computational problem, based on previous work in [7, 19]. In Section 6.2 we use a polar representation to choose an appropriate basis, and show that the relevant inner products can be computed, as in Section 5.2. Finally, in Section 6.3 we string the preceding results together and obtain the Σ_1-classification.

6.1 The Dirac operator in two dimensions

Recently, the authors of [7] studied the addition of a singular Coulomb-type potential to the free Dirac operator described above. By using perturbation theory, they find that the essential spectrum (cf. Definition 2.17) coincides with that of the free Dirac operator, i.e., $(-\infty, 1] \cup [1, \infty)$. However, they are unable to say anything about the discrete spectrum of the operator, other than that it must obviously be contained in the interval $(-1, 1)$. The goal of this chapter is to demonstrate a Σ_1-classification for the operator discussed in [7], and hence provide a numerical algorithm that for any given level of prevision $\delta > 0$ will eventually find every point in $(-1, 1)$ which is within δ of the discrete spectrum.

Mathematical setup and previous work

We begin with defining the two-dimensional Dirac operator with boundary conditions, following [7]. The free Dirac operator acting on the underlying Hilbert space $\mathcal{H} \coloneqq L^2(\mathbb{R}^2)$ is formally defined by

$$H_0 \coloneqq -i\boldsymbol{\sigma} \cdot \nabla + m\sigma_3 = \begin{pmatrix} 1 & -i(\partial_{x_1} - i\partial_{x_2}) \\ -i(\partial_{x_1} + \partial_{x_2})) & -1 \end{pmatrix},$$

where $m \geq 0$ is the mass of the particle and $\boldsymbol{\sigma} \coloneqq (\sigma_1, \sigma_2)$, where we recall the *Pauli matrices* $\sigma_i \in \mathbb{C}^{2\times 2}$

$$\sigma_1 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_2 = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_3 = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}.$$

As in Chapter 5, the operator ∇ is defined by $\nabla \coloneqq (\nabla_1, \nabla_2)$ where $\nabla_i = (\partial_i, \partial_i)^T$ for $i = 1, 2$ and $\boldsymbol{\sigma} \cdot \nabla \coloneqq \sigma_1\nabla_1 + \sigma_2\nabla_2$. Going forward we will set $m = 1$ to simplify notation since this will not have any impact on the discussion. However the operator σ_3 in H_0 will sometimes be referred to as the *mass term*, and we note in particular that this forms a bounded self-adjoint (and hence everywhere defined) operator in $L^2(\mathbb{R}^2)$. As remarked in [7], H_0 is self-adjoint on the first Sobolev space $H^1(\mathbb{R}^2; \mathbb{C}^2)$ with a purely essential spectrum $\mathrm{Sp}(H_0) = (-\infty, -1] \cup [1, +\infty)$. This can be shown with a completely analogous proof to that of [31, Theorem 1.1] in the standard three dimensional case.

Next, we need to formulate the *infinite mass boundary conditions*. Let $\Omega \subset \mathbb{R}^2$ be a connected domain with a regular boundary $\partial\Omega$. We then denote the outward normal vector by $\boldsymbol{n}$ and the tangent vector by $\boldsymbol{t}$ so that $(\boldsymbol{n}, \boldsymbol{t})$ gives a positive orientation of $\partial\Omega$. The infinite mass boundary conditions are defined by

$$\mathbf{B}_{\boldsymbol{n}}\psi = \psi \text{ on } \partial\Omega,$$

where the *boundary matrix* is defined

$$\mathbf{B}_{\boldsymbol{n}} \coloneqq -i\boldsymbol{\sigma} \cdot \boldsymbol{n}.$$

In our case, the domain of interest is the open sector of aperture $\omega \in (0, 2\pi]$ centered at the origin:

$$S_\omega \coloneqq \{(r\cos\theta, r\sin\theta) \in \mathbb{R}^2 : r > 0, 0 < \theta < \omega\}. \tag{6.1}$$

The boundary of S_ω is just the union of the x-axis $\{(r, 0) : r \geq 0\}$ and the ray at angle ω out from the origin, $\{(r, \omega) : r \geq 0\}$. These two cases give $\boldsymbol{n} = (0, -1)$ and $\boldsymbol{n} = (-\sin\omega, \cos\omega)$ respectively. By straightforward computation the boundary matrices then become

$$\mathbf{B}_{\boldsymbol{n}} = -i\sigma_3(-\sigma_2) = i\sigma_3\sigma_2 = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} = \sigma_1, \quad \boldsymbol{n} = (0, -1),$$

$$\mathbf{B}_{\boldsymbol{n}} = -i\sigma_3(\sigma_1(-\sin\omega) + \sigma_2\cos\omega) = \begin{pmatrix} 0 & -\mathrm{e}^{-i\omega} \\ -\mathrm{e}^{i\omega} & 0 \end{pmatrix}, \quad \boldsymbol{n} = (-\sin\omega, \cos\omega).$$

It follows $u = (u_1, u_2)^T \in \mathcal{H}$ satisfies the boundary conditions if:

$$u_1(r, 0) = u_2(r, 0) \qquad \forall r \geq 0, \tag{6.2}$$

$$u_1(r, \omega) = u_2(r, \omega)(-\mathrm{e}^{-i\omega}) \quad \forall r \geq 0. \tag{6.3}$$

Going forward, we will just use the notation $\mathbf{B}_{\boldsymbol{n}}u = u$ to mean that u satisfies both boundary conditions (6.2) and (6.3).

The topic of self-adjointness for the free Dirac operator on S_ω was studied in [19, Theorem 1.2] which showed that this is determined by the convexity of the domain:

Theorem 6.1. *Let $\omega \in (0, 2\pi]$, S_ω as in (6.1), and let H_ω be the operator*

$$\begin{aligned} &H_\omega \psi \coloneqq H_0 \psi, \\ &\mathcal{D}(H) \coloneqq \{u \in H^1(S_\omega; \mathbb{C}^2) : \mathbf{B}_{\boldsymbol{n}} u = u\} \end{aligned} \tag{6.4}$$

Then:

(i) *If $0 < \omega \leq \pi$, then H_ω is self-adjoint.*

(ii) *If $\pi < \omega \leq 2\pi$ then H_ω has infinitely many self-adjoint extensions with a unique "distinguished" one whose domain is contained in the Sobolev space $H^{1/2}(S_\omega; \mathbb{C}^2)$.*

Remark 6.2. Without going into the theory of Sobolev spaces and partial differential equations, which is outside the scope of this thesis, it seems strange to impose boundary conditions on a set of measure zero in $L^2(\mathbb{R}^2)$. However, as alluded to in Chapter 2, the regularity of the domain H^1 ensures that this can be given a well defined meaning via the so-called *trace operator*, see e.g. [12].

In [7] the object of study is the perturbation of the operator H_ω defined in (6.4) by a Coulomb-potential V defined as (note that S_ω does not contain the origin):

$$V(x) \coloneqq \frac{\nu}{|x|} \mathbf{1}_2 \text{ for all } x \in S_\omega, \tag{6.5}$$

where $\mathbf{1}_2$ is the 2×2 identity matrix and $\nu > 0$. More precisely, they first define the *minimal* operator $H_{\min}$ by:

$$\begin{aligned} &H_{\min} \psi \coloneqq (H_0 + V)u, \\ &\mathcal{D}(H_{\min}) \coloneqq \{u \in C_c^\infty(\overline{S_\omega} \setminus \{0\}; \mathbb{C}^2) : \mathbf{B}_{\boldsymbol{n}} u = u\}, \end{aligned} \tag{6.6}$$

where $C_c^\infty(\overline{S_\omega} \setminus \{0\}; \mathbb{C}^2)$ is the set of all functions $g = (g_1, g_2)^T$ such that g_i is an infinitely differentiable complex valued function with compact support contained in $\overline{S_\omega} \setminus \{0\}$ for $i = 1, 2$. Note that the origin is *not* inside the support of the g_i. We have the following two key results [7, Theorems 1.7 and 1.9]:

Theorem 6.3. *Let $\omega \in (0, 2\pi]$, S_ω as in (6.1) and H_{min} as in (6.6). Then*

(i) *If $\nu^2 \leq \frac{\pi^2 - \omega^2}{4\omega^2}$, then H_{min} is essentially self-adjoint and its self-adjoint closure is given by*

$$\mathcal{D}(\overline{H_{min}}) = \mathcal{D}(H_\omega) = \{u \in H^1(S_\omega; \mathbb{C}^2) : \mathbf{B}_{\boldsymbol{n}} u = u\},$$

(ii) *If $\nu^2 > \frac{\pi^2 - \omega^2}{4\omega^2}$ then H_{min} has infinitely many self-adjoint extensions.*

Theorem 6.4. *Let H_{min} be as in (6.6) and let T be any self-adjoint extension of H_{min}. Then T has essential spectrum (see Definition 2.17)*

$$\mathrm{Sp}_{\mathrm{ess}}(T) = (-\infty, -1] \cup [1, +\infty).$$

Without the boundary conditions, the so-called *partial wave subspace decomposition* can be used to show that the entire spectrum of the Dirac-Coulomb operator is as in Theorem 6.4. This is a common idea when dealing with spherically symmetric potentials that provides a diagonalisation of the Dirac

operator in terms of operators in the radial component only. See for example [31, Chapter 4.6] for the three dimensional case, which would work equally well in two dimensions. In [7] this approach is combined with perturbation theory in order to prove Theorem 6.4 for the Dirac-Coulomb operator with the infinite mass boundary conditions and characterise the essential spectrum. As they point out, however, the boundary conditions prevent this approach from providing any information about the discrete spectrum, other than that it must be contained in $(-1, 1)$.

Thus our goal is to give a Σ_1-result which would yield a numerical algorithm with rigorous error control on how far away from the true spectrum its output is. We will focus our attention to the (self-adjoint) closure of the operators covered by Theorem 6.3 (i). In the language of the SCI hierarchy, the domain of our computational problem is

$$\Omega \coloneqq \left\{ \overline{H_{\min}} : H_{\min} \text{ is as in Theorem 6.3 (i)} \right\}. \tag{6.7}$$

For each $H \in \Omega$ we have $\mathcal{D}(H) = \mathcal{D}(H_\omega)$ as in Theorem 6.3. With the domain of the computational problem set up, we can start to build the algorithm, following the same overall strategy as in Chapter 5.

6.2 Approximating the inverse resolvent norm from matrix evaluations

In the following, we let H denote an arbitrary operator in the computational domain Ω as defined in (6.7). We will follow the same strategy as in Chapter 5, using Lemma 3.25 and Lemma 5.3 to compute approximations to the reciprocal of the resolvent norm. This in turn will let us apply Proposition 3.29. Thus the two key objectives in building the Σ_1-algorithm are:

1. Find an orthonormal basis $\{\psi_j\}_{j=1}^{\infty} \subset \mathcal{D}(H)$ of $\mathcal{H} = L^2(S_\omega; \mathbb{C}^2)$ whose linear span forms a core for H.
2. Show that the inner products $\langle (H - zI)\psi_m, (H - zI)\psi_n \rangle$ can be computed to arbitrary precision using finitely many arithmetic operations and comparisons.

The polar Dirac operator

The idea behind our approach is to decompose the space $\mathcal{H} = L^2(S_\omega; \mathbb{C}^2)$ into a radial and an angular part, see also [7, pp. 11–12]. We use the standard polar coordinates in the plane:

$$\begin{aligned} x_1 &= r\cos\theta, \quad r \coloneqq \sqrt{x_1^2 + x_2^2} \in (0, +\infty), \\ x_2 &= r\sin\theta, \quad \theta \coloneqq \operatorname{sign}(x_2)\cos^{-1}(x_1/r) \in [0, 2\pi). \end{aligned}$$

For any $\psi \in L^2(S_\omega; \mathbb{C})$, consider the map $\psi \mapsto \sqrt{r}\psi =: \varphi$. We view φ as an element of the product space $L^2((0,\infty) \times (0,\omega), \mathrm{d}r \otimes \mathrm{d}\theta)$. Since the Jacobian when changing to polar coordinates is r, we have

$$\|\psi\|^2 = \int_{S_\omega} |\psi(x)|^2 \,\mathrm{d}x = \int_0^\omega \int_0^\infty |\varphi(r,\theta)|^2 \,\mathrm{d}r\,\mathrm{d}\theta = \|\varphi\|^2.$$

Clearly the map $\psi \mapsto \varphi$ is bijective so we have a unitary Hilbert space isomorphism

$$L^2(S_\omega;\mathbb{C}) \simeq L^2((0,\infty)\times(0,\omega),\mathrm{d}r\otimes\mathrm{d}\theta) \simeq L^2((0,\infty),\mathrm{d}r)\otimes L^2((0,\omega),\mathrm{d}\theta).$$

Reasoning like this for both components of $\psi \in L^2(S_\omega;\mathbb{C}^2)$, we get the isometric isomorphism

$$L^2(S_\omega;\mathbb{C}^2) \simeq L^2((0,\infty),\mathrm{d}r)\otimes L^2((0,\omega);\mathbb{C}^2) =: \mathcal{H}'.$$

Next we will represent the Dirac operator acting on the Hilbert space $\mathcal{H}'$ using polar coordinates. To do this, we define the polar unit vectors

$$\boldsymbol{e}_r \coloneqq (\cos\theta,\sin\theta), \quad \boldsymbol{e}_\theta \coloneqq (-\sin\theta,\cos\theta).$$

Recall that the gradient in polar coordinates is

$$\nabla = \partial_r \boldsymbol{e}_r + \frac{1}{r}\partial_\theta \boldsymbol{e}_\theta,$$

where we use the shorthand

$$\partial_r \coloneqq \frac{\partial}{\partial r}, \quad \partial_\theta \coloneqq \frac{\partial}{\partial\theta}.$$

By straightforward computation, we have the relations

$$\boldsymbol{\sigma}\cdot\boldsymbol{e}_r = \begin{pmatrix} 0 & e^{-i\theta} \\ e^{i\theta} & 0 \end{pmatrix}, \quad \boldsymbol{\sigma}\cdot\boldsymbol{e}_\theta = i\boldsymbol{\sigma}\cdot\boldsymbol{e}_r\sigma_3.$$

This gives

$$-i\boldsymbol{\sigma}\cdot\nabla = -i\boldsymbol{\sigma}\cdot\left(\partial_r\boldsymbol{e}_r + \frac{1}{r}\partial_\theta\boldsymbol{e}_\theta,\right) = -i\boldsymbol{\sigma}\cdot\boldsymbol{e}_r\left(\partial_r + \frac{1}{2r} - \frac{1}{r}K_\omega\right),$$

where

$$K_\omega \coloneqq \frac{1}{2}\mathbf{1}_2 - i\sigma_3\partial_\theta$$

is the so-called *spin-orbit* operator, to which we will return shortly. Thus the polar Dirac operator acting on $\mathcal{H}'$ reads as

$$H = -i\boldsymbol{\sigma}\cdot\boldsymbol{e}_r\left(\partial_r + \frac{1}{2r} - \frac{1}{r}K_\omega\right) + \sigma_3 + \frac{\nu}{r}\mathbf{1}_2 \tag{6.8}$$

We will use this equivalent representation of the Dirac operator without going forward. For the standard Dirac-Coulomb operator without boundary conditions, the properties of K_ω yield a decomposition into a direct sum of one dimensional Dirac operators acting on $L^2(0,\infty)$. This is not possible for the Dirac operator with the mass term when boundary conditions are imposed, as pointed out in [7]. However, as we will see, we can still give a Σ_1-algorithm for the spectrum.

Choice of basis

The next step required is to pick an orthonormal basis for $\mathcal{H}' = L^2((0,\infty),\mathrm{d}r) \otimes L^2((0,\omega);\mathbb{C}^2)$ that satisfies the two points listed at the start of this section. Of course, the basis functions need to satisfy the boundary conditions. As discussed in Section 2.3, a basis for $\mathcal{H}'$ can be constructed by finding basis functions for $L^2(0,\infty)$ and $L^2((0,\omega);\mathbb{C}^2)$ and then taking the product of these two bases. We begin with the space $L^2((0,\omega);\mathbb{C}^2)$ where we have the following result from [7, p. 12]:

Proposition 6.5. *Let $\omega \in (0,2\pi]$, S_ω as in (6.1) and set $\lambda_k \coloneqq \frac{(2k+1)\pi}{2\omega}$ for $k \in \mathbb{N}$. For $k \in \mathbb{N}$, we define the following functions in $L^2((0,\omega);\mathbb{C}^2)$:*

$$f_k^+(\theta) \coloneqq \frac{1}{\sqrt{2\omega}} \begin{pmatrix} e^{+i(\lambda_k - \frac{1}{2})\theta} \\ e^{-i(\lambda_k - \frac{1}{2})\theta} \end{pmatrix}, \quad f_k^-(\theta) \coloneqq \frac{-i}{\sqrt{2\omega}} \begin{pmatrix} e^{-i(\lambda_k + \frac{1}{2})\theta} \\ e^{+i(\lambda_k + \frac{1}{2})\theta} \end{pmatrix}.$$

Then spin-orbit operator $K_\omega = \frac{1}{2}\mathbf{1}_2 - i\sigma_3\partial_\theta$ is self-adjoint on

$$\mathcal{D}(K_\omega) \coloneqq \{\phi = (\phi_1,\phi_2) \in H^1((0,\omega);\mathbb{C}^2) : \phi_1(0) = \phi_2(0), \phi_2(\omega) = e^{-i\omega}\phi_1(\omega)\},$$

with the additional properties that

(i) *$\{f_k^\pm\}_{k\in\mathbb{N}}$ is an orthonormal basis for $L^2((0,\omega);\mathbb{C}^2)$ of eigenfunctions of K_ω with eigenvalues $\{\pm\lambda_k\}_{k\in\mathbb{N}}$*

(ii) *$-i(\boldsymbol{\sigma}\cdot\boldsymbol{e}_r)f_k^\pm = \pm f_k^\mp$.*

We enumerate the basis $\{f_k^\pm\}_{k\in\mathbb{N}}$ as $\{f_k\}_{k\in\mathbb{Z}_{\geq 0}}$ where

$$f_{2k} \coloneqq f_k^+, \quad f_{2k+1} \coloneqq f_k^-, \quad k \geq 0.$$

Next, we need a basis for the radial component $L^2((0,\infty),\mathrm{d}r)$. We will base this on the *generalised Laguerre polynomials*, found in e.g. [29, Chapter 5]. For any real number $\alpha > -1$, the family of Laguerre polynomials $\{L_n^\alpha\}_{n=0}^\infty$ are the solutions to the second order differential equation,

$$xy'' - (\alpha + 1 - x)y' + ny = 0, \quad n \geq 0. \tag{6.9}$$

They form an orthogonal set in the weighted space $L^2((0,\infty), x^\alpha e^{-x}\,\mathrm{d}x)$, satisfying

$$\int_0^\infty e^{-x}x^\alpha L_n^\alpha(x)L_m^\alpha(x)\,\mathrm{d}x = \frac{\Gamma(n+\alpha+1)}{n!}\delta_{nm}, \tag{6.10}$$

where Γ is the gamma function, $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$. We also have the recurrence relations

$$L_n^{\alpha+1}(x) = \sum_{i=0}^n L_i^\alpha, \tag{6.11}$$

$$(L_n^\alpha)' = -L_{n-1}^{\alpha+1}, \tag{6.12}$$

$$xL_{n-1}^{\alpha+1} = (n+\alpha)L_{n-1}^\alpha - nL_n^\alpha, \tag{6.13}$$

Furthermore, as shown in [29], the set of functions $\{x^{\alpha/2}e^{-x/2}L_n^\alpha(x)\}_{n=1}^\infty$ is a complete orthogonal set in $L^2(0,\infty)$. In view of this fact and (6.10), for $\alpha \in \mathbb{N}$, the *generalised Laguerre functions* $\{h_n^\alpha\}_{n\geq 0}$ defined by

$$h_n^\alpha(x) \coloneqq N_n^\alpha x^{\alpha/2}e^{-x/2}L_n^\alpha(x), \quad x \geq 0, \tag{6.14}$$

where $N_n^\alpha = \sqrt{n!/(n+\alpha)!}$, form an orthonormal basis for $L^2(0,\infty)$. In order to ensure that the basis functions work nicely with the Coulomb potential we will use $\alpha = 2$ and so the basis we will use is defined

$$h_n(r) \coloneqq N_n r e^{-r/2} L_n^2(r), \tag{6.15}$$

where $N_n = [(n+1)(n+2)]^{-1/2}$. Thus an orthonormal basis for $\mathcal{H}'$ is given by the products

$$\{h_n f_k\}_{n,k\geq 0} = \{h_{m_1} f_{m_2} \colon (m_1, m_2) \in \mathbb{Z}^2_{\geq 0}\}. \tag{6.16}$$

Note that since the functions $\{f_k\}$ satisfy the conditions given in the definition of $\mathcal{D}(K_\omega)$ in Proposition 6.5, the functions in (6.16) satisfy the boundary conditions (6.2) and (6.3). In order to enumerate these products set over $\mathbb{N}$ we may do as in Chapter 5 and consider the sets $S_n \coloneqq \{h_{m_1} f_{m_2} : |m| = m_1 + m_2 \leq n\}$. We have

$$\begin{aligned} S_0 &= \{h_0 f_0\}, \\ S_1 &= S_0 \cup \{h_1 f_0, h_0 f_1\}, \\ S_2 &= S_1 \cup \{h_2 f_0, h_1 f_1, h_0 f_2\}, \end{aligned}$$

and so on. If we let r_n denote the number of elements in S_n then it is easily seen that $r_n = \sum_{i=0}^n (i+1) = \frac{1}{2}(n+2)(n+1)$. We label $h_0 f_0$ as ψ_1 and the elements of $S_1 \setminus S_0$ as $\{\psi_2, \psi_3\}$. Again, the particular ordering within each successive "half-sphere" is not important as long as we fix one. In general, given the enumeration $\{\psi_1, \ldots, \psi_{r_n}\}$ of S_n we enumerate the $r_{n+1} - r_n = n+2$ elements of $S_{n+1} \setminus S_n$ as $\{\psi_{r_n+1}, \ldots, \psi_{r_{n+1}}\}$. By proceeding inductively we obtain an enumeration of the functions in (6.16) which we denote by $\mathcal{B}_{\mathcal{H}'} \coloneqq \{\psi_n\}_{n=1}^\infty$. This enumeration defines a computable bijective map $\zeta : \mathbb{Z}^2_{\geq 0} \to \mathbb{N}$ whose inverse we denote by $\xi : \mathbb{N} \to \mathbb{Z}^2_{\geq 0}$ so that for each $m = (m_1, m_2) \in \mathbb{Z}^2_{\geq 0}$ we have $h_{m_1} f_{m_2} = \psi_{\zeta(m)}$ and for each $n \in \mathbb{N}$ we have $\psi_n = h_{\zeta(n)_1} f_{\zeta(n)_2}$, writing $\xi(n)_1$ and $\xi(n)_2$ for the first and second components of $\xi(n)$, respectively. Also note that $|\xi(n)| \leq n$, and hence in particular $\xi(n)_1$ grows with n at a sub-linear rate. By construction, the fastest growing subsequence of $\{\xi(n)_1\}_n$ grows like $\sqrt{n}$. Finally we need:

Lemma 6.6. *Let $H \in \Omega$ be as in (6.7) and let $\mathcal{B}_{\mathcal{H}'} = \{\psi_n\}_{n=1}^\infty$ be the basis constructed above. Then $\mathcal{S} \coloneqq \operatorname{Span}\{\psi_n\}_{n=1}^\infty$ forms a core for H.*

Proof. We will follow the same strategy as in the proof of Lemma 5.5. Since the "mass term" σ_3 is a bounded self-adjoint operator we can omit it from the formal definition of H for the purposes of this proof in order to simplify notation. We know from Theorem 6.3 (i) that $\mathcal{C} \coloneqq \left\{g \in C_c^\infty(\overline{S_\omega} \setminus \{0\}; \mathbb{C}^2) : \mathbf{B}_{\boldsymbol{n}} g = g\right\}$ is a core for H and that $\mathcal{S} \subset \mathcal{D}(H)$. Let T denote the closure of the formal operator $-i\boldsymbol{\sigma} \cdot \nabla + \frac{\nu}{r}\mathbf{1}_2$ restricted to the domain $\mathcal{S}$. We will show that $\mathcal{C} \subset \mathcal{D}(T)$. Then $H \restriction \mathcal{C} \subset T \subset H$, giving $H = \overline{H \restriction \mathcal{C}} \subset T \subset H$ so that $T = H$ and hence $\mathcal{S}$ is a core for H. To show the inclusion $\mathcal{C} \in \mathcal{D}(T)$ we let $g \in \mathcal{C}$ and consider the sequence $\{g_m\}_{m=1}^\infty \subset \mathcal{S}$ defined by

$$g_m = \sum_{n=1}^m \langle g, \psi_n \rangle \psi_n$$

which converges to g in $\mathcal{H}'$. If the sequence Tg_m also converges in $\mathcal{H}'$ then it follows that $g \in \mathcal{D}(T)$ since T is a closed operator by assumption. Now

$$Tg_m = \sum_{n=1}^{m} \langle g, \psi_n \rangle T\psi_n = -i \sum_{n=1}^{m} \langle g, \psi_n \rangle \left(-i\boldsymbol{\sigma} \cdot \nabla + \frac{\nu}{r} \mathbf{1}_2 \right) \psi_n,$$

so we see that Tg_m converges if the positive series

$$\sum_{n=1}^{\infty} |\langle g, \psi_n \rangle| \left\| \left(-i\boldsymbol{\sigma} \cdot \nabla + \frac{\nu}{r} \mathbf{1}_2 \right) \psi_n \right\|$$

converges in $\mathbb{R}$. Once again, we will prove this in two parts by showing that:

1. The sequence $|\langle g, \psi_n \rangle|$ tends to zero faster than the reciprocal of any polynomial in n.
2. The norms $\| \left(-i\boldsymbol{\sigma} \cdot \nabla + \frac{\nu}{r} \mathbf{1}_2 \right) \psi_n \|$ grow only polynomially in n.

In the proof for the Hermite functions in Chapter 5, we used a trick involving the well known Harmonic oscillator to show the first point. With Laguerre functions, we need to look for another operator that does the same trick. In [11], symmetry properties of the generalised Laguerre functions in (6.14) are studied. Here, the author uses the differential equation (6.9) and recurrence relations to derive so-called "ladder operators" acting on $L^2(0, \infty)$,

$$\begin{aligned} \tilde{\mathcal{L}}^+ &\coloneqq r \frac{\mathrm{d}}{\mathrm{d}r} - \frac{r}{2} + \frac{\alpha}{2} + n + 1, \\ \tilde{\mathcal{L}}^- &\coloneqq -r \frac{\mathrm{d}}{\mathrm{d}r} - \frac{r}{2} + \frac{\alpha}{2} + n. \end{aligned}$$

Then they define the Hamiltonian $\tilde{D} = \tilde{\mathcal{L}}^+ \tilde{\mathcal{L}}^-$ which satisfies [11, Equation (23)]

$$\tilde{D} h_n^\alpha = n(n + \alpha) h_n^\alpha.$$

In our case with $\alpha = 2$ we define the operator D acting component-wise on $u \in \mathcal{H}'$ by

$$\begin{aligned} Du &= \mathcal{L}^+ \mathcal{L}^- u, \\ \mathcal{L}^+ &\coloneqq r \frac{\partial}{\partial r} - \frac{r}{2} + n + 2, \\ \mathcal{L}^- &\coloneqq -r \frac{\partial}{\partial r} - \frac{r}{2} + n + 1. \end{aligned}$$

Then for any $\psi_n \in \mathcal{B}_{\mathcal{H}'}$, D acts non-trivially only on the radial component, i.e., given $\psi_n = h_{\xi(n)_1} f_{\xi(n)_2}$ we have

$$D\psi_n = \left(\tilde{D} h_{\xi(n)_1} \right) f_{\xi(n)_2} = \xi(n)_1 (\xi(n)_1 + 2) \psi_n.$$

By assumption, g is infinitely differentiable with compact support. Hence $g \in \mathcal{D}(D^k)$ for any $k \in \mathbb{N}$, and

$$\langle D^k g, \psi_n \rangle = \langle g, D^k \psi_n \rangle = \left[\xi(n)_1 (\xi(n)_1 + 2) \right]^k \langle g, \psi_n \rangle.$$

Thus the sequence $\left\{[\xi(n)_1(\xi(n)_1+2)]^k\,|\langle g,\psi_n\rangle|\right\}_{n\geq 1}$ must be square summable in $\mathbb{R}$. Since $k\in\mathbb{N}$ can be arbitrarily large, this shows that the sequence $|\langle g,\psi_n\rangle|$ must go to zero faster than the reciprocal of any polynomial in n. To prove the second part of the claim, we consider how the operator $-i\boldsymbol{\sigma}\cdot\nabla+\frac{\nu}{r}\mathbf{1}_2$ acts on any given ψ_n. So let $n\in\mathbb{N}$ be given and simplify notation by setting $(m,k):=\xi(n)$ so that $\psi_n=h_mf_k$. Recalling from Proposition 6.5 that $K_\omega f_j^\pm=\pm\lambda_k f_j^\pm$ and $-i(\boldsymbol{\sigma}\cdot\boldsymbol{e}_r)f_j^\pm=\pm f_j^\mp$ we have

$$\begin{aligned}
H\psi_n(r,\theta)&=\left[-i\boldsymbol{\sigma}\cdot\boldsymbol{e}_r\left(\partial_r+\frac{1}{2r}-\frac{1}{r}K_\omega\right)+\frac{\nu}{r}\mathbf{1}_2\right]h_m(r)f_k(\theta)\\
&=-i\boldsymbol{\sigma}\cdot\boldsymbol{e}_r\left(h_m'(r)f_k(\theta)+\frac{h_m(r)}{2r}f_k(\theta)-\frac{h_m(r)}{r}\lambda_k f_k(\theta)\right)\\
&\quad+\frac{\nu}{r}h_m(r)f_k(\theta).\\
&=\left(h_m'(r)-\left(\lambda_k-\frac{1}{2}\right)\frac{h_m(r)}{r}\right)(-i\boldsymbol{\sigma}\cdot\boldsymbol{e}_r)f_k(\theta)+\frac{\nu}{r}h_m(r)f_k(\theta).
\end{aligned}$$

Hence by orthonormality of the set $\{f_j^\pm\}_j$ in $L^2((0,\omega);\mathbb{C}^2)$ we have

$$\begin{aligned}
\|H\psi_n\|_{\mathcal{H}'}&\leq\left\|\left(h_m'-\left(\lambda_k-\frac{1}{2}\right)\frac{h_m}{r}\right)f_k\right\|_{\mathcal{H}'}+\left\|\frac{\nu}{r}h_mf_k\right\|_{\mathcal{H}'}\\
&=\left\|h_m'-\left(\lambda_k-\frac{1}{2}\right)\frac{h_m}{r}\right\|_{L^2(0,\infty)}+\nu\left\|\frac{h_m}{r}\right\|_{L^2(0,\infty)}.
\end{aligned}$$

Now, for the Laguerre functions in (6.15) we have

$$h_m'=\left(\frac{1}{r}-\frac{1}{2}\right)h_m+N_mre^{-r/2}(L_m^2)'. \tag{6.17}$$

Thus,

$$\|H\psi_n\|_{\mathcal{H}'}\leq N_m\left\|re^{-r/2}(L_m^2)'\right\|_{L^2(0,\infty)}+\left(\left|\lambda_k-\frac{3}{2}\right|+\nu\right)\left\|\frac{h_m}{r}\right\|_{L^2(0,\infty)}+\frac{1}{2}.$$

By construction m and k, and hence also λ_k, are implicitly functions of n that grow sub-linearly in n. So the only thing left to check is that the terms

$$N_m\left\|re^{-r/2}(L_m^2)'\right\|_{L^2(0,\infty)}\quad\text{and}\quad\left\|\frac{h_m}{r}\right\|_{L^2(0,\infty)} \tag{6.18}$$

grow polynomially in m. For the first term, the recurrence relation (6.12) and the identity (6.10) give

$$\begin{aligned}
\left\|re^{-r/2}(L_m^2)'\right\|_{L^2(0,\infty)}&=\int_0^\infty r^2e^{-r}\left((L_m^2(r))'\right)^2\,\mathrm{d}r\\
&=\int_0^\infty r^2e^{-r}\left(L_{m-1}^{(3)}(r)\right)^2\,\mathrm{d}r\\
&\leq\int_0^\infty r^3e^{-r}\left(L_{m-1}^{(3)}(r)\right)^2\,\mathrm{d}r
\end{aligned}$$

$$= \frac{\Gamma(m-1+3+1)}{\Gamma(m-1)}$$
$$= (m+2)(m+1)m(m-1).$$

Since $N_m = [(m+1)(m+2)]^{-1/2}$ we can conclude that the first term of (6.18) grows polynomially in m. For the second term we have :

$$\left\| \frac{h_m}{r} \right\|^2_{L^2(0,\infty)} = \int_0^\infty e^{-r} \left(L_m^2(r) \right)^2 \mathrm{d}r. \tag{6.19}$$

By the recurrence relation (6.11),

$$L_m^2 = \sum_{i=0}^{m} \sum_{j=0}^{i} L_j = \sum_{i=0}^{m} (m-(i-1)) L_i, \tag{6.20}$$

where $L_i \coloneqq L_i^0$. By orthogonality of the L_m with respect to to the weight e^{-r}, when squaring the sum L_m^2 in (6.19) we can drop all the cross terms so that:

$$\begin{aligned}
\left\| \frac{h_m}{r} \right\|^2_{L^2(0,\infty)} &= \int_0^\infty e^{-r} \sum_{i=0}^{m} (m-(i-1))^2 L_i(r)^2 \,\mathrm{d}r \\
&= \sum_{i=0}^{m} (m-(i-1))^2 \int_0^\infty e^{-r} L_i(r)^2 \,\mathrm{d}r \\
&= \sum_{i=0}^{m} (m-(i-1))^2 \cdot 1 \\
&= \frac{m+1}{6}(2m^2+7m+6).
\end{aligned}$$

Thus we have shown both parts of the above claim and the lemma follows. ■

Computing inner products

Having established that $\mathcal{B}_{\mathcal{H}'} = \{\psi_n\}_{n=1}^\infty$ is an orthonormal basis whose span is a core for $H = -i\boldsymbol{\sigma} \cdot \nabla + \sigma_3 + \frac{\nu}{r}\mathbf{1}_2$ and satisfies the boundary conditions, the final thing to prove is that the inner products

$$\langle (H - zI)\psi_m, (H - zI)\psi_n \rangle \tag{6.21}$$

can be computed to arbitrary precision for any $m, n \in \mathbb{N}$ using finitely many arithmetic operations and comparisons. The additive term σ_3 in H only changes the wavefunction by multiplying the second component by -1 so this term does not affect the computability of the inner products above. Hence we drop the mass term in the formal definition of H for this part of the discussion as well. First we make the general observation that by definition of the inner product in $\mathcal{H}'$, for arbitrary $h, g \in L^2(0,\infty)$ and for $f_i, f_j \in \{f_k^\pm\}_k \subset L^2((0,\omega);\mathbb{C}^2)$ we have

$$\langle hf_i, gf_j \rangle_{\mathcal{H}'} = \langle h, g \rangle_{L^2(0,\infty)} \cdot \langle f_i, f_j \rangle_{L^2((0,\omega);\mathbb{C}^2)} = \langle h, g \rangle_{L^2(0,\infty)} \cdot \delta_{ij} \tag{6.22}$$

By our choice of basis $\mathcal{B}_{\mathcal{H}'}$, every basis function is on the form $h_m f_k^+$ or $h_m f_k^-$ for some $m, k \in \mathbb{N}$. We will show how to compute (6.21) for $\psi_n = h_{n_1} f_{n_2}^+$

and $\psi_m = h_{m_1} f^+_{m_2}$ where $n, m \in \mathbb{Z}^2_{\geq 0}$ (we are slightly abusing notation and writing (n_1, n_2) for $\xi(n)$ and likewise for m). The other possible cases are argued in the exact same way. We calculate as in the proof of Lemma 6.6 with $H = -i\boldsymbol{\sigma} \cdot \nabla + \frac{\nu}{r}\mathbf{1}_2$,

$$(H - zI)\psi_n = \left(h'_{n_1} - \left(\lambda_{n_2} - \frac{1}{2}\right)\frac{h_{n_1}}{r}\right)(-i(\boldsymbol{\sigma}\cdot\boldsymbol{e}_r)f_{n_2}) + \left(\frac{\nu}{r} - z\right)h_{n_1} f_{n_2},$$

and similarly for $(H - zI)\psi_m$. Using the observation (6.22), and the fact that $-i(\boldsymbol{\sigma}\cdot\boldsymbol{e}_r)f^{\pm}_j = \pm f^{\mp}_j$, it is sufficient to show that for indices $m, n \in \mathbb{N}$, inner products in $L^2(0,\infty)$ on one of the three forms

$$\begin{aligned}
&\left\langle h'_n - \left(\lambda_n - \frac{1}{2}\right)\frac{h_n}{r}, h'_m - \left(\lambda_m - \frac{1}{2}\right)\frac{h_m}{r}\right\rangle,\\
&\left\langle \left(\frac{1}{r} - 1\right)h_n, \left(\frac{1}{r} - 1\right)h_m\right\rangle,\\
&\left\langle h'_n - \left(\lambda_n - \frac{1}{2}\right)\frac{h_n}{r}, \left(\frac{1}{r} - 1\right)h_m\right\rangle,
\end{aligned} \tag{6.23}$$

where $\lambda_n, \lambda_m \in \mathbb{Q}$ are given constants, can be computed to arbitrary precision. First note that by (6.17) we have

$$h'_n - \left(\lambda_n - \frac{1}{2}\right)\frac{h_n}{r} = -\frac{1}{2}h_n + \left(\frac{3}{2} - \lambda_n\right)\frac{h_n}{r} + N_n r e^{-r/2}(L^2_m)'. \tag{6.24}$$

From this we see that the second and third forms of inner products in (6.23) are contained in the calculation of the first one, so we can restrict our attention to that form without loss of generality. In the last term of (6.24), we use the recurrence relations (6.13) and (6.12) to obtain (recall $L_j \coloneqq L^0_j$),

$$r(L^2_n)' = nL^2_n - (n+2)L^2_{n-1} = n\sum_{i=0}^{n}\sum_{j=0}^{i} L_j - (n+2)\sum_{i=0}^{n-1}\sum_{j=0}^{i} L_j$$

Hence by linearity of the inner product, writing $r(L^2_n)'$ and $r(L^2_m)'$ as sums of polynomials L_j, to compute the first inner product in (6.23) is a finite linear combination with computable coefficients of inner products,

$$\left\langle h_n + \frac{h_n}{r} + e^{-r/2}L_i, h_m + \frac{h_m}{r} + e^{-r/2}L_j\right\rangle, \tag{6.25}$$

where $i, j \in \mathbb{N}$ are arbitrary. Expanding (6.25) we see that the first type of inner product in (6.23) (and hence all three) is a linear combination with known coefficients of inner products on one of the forms

(a) $\langle h_n, h_m\rangle$ (b) $\langle h_n, \frac{h_m}{r}\rangle$ (c) $\langle h_n, e^{-r/2}L_j\rangle$

(d) $\langle \frac{h_n}{r}, h_m\rangle$ (e) $\langle \frac{h_n}{r}, \frac{h_m}{r}\rangle$ (f) $\langle \frac{h_n}{r}, e^{-r/2}L_j\rangle$

(g) $\langle e^{-r/2}L_i, h_m\rangle$ (h) $\langle e^{-r/2}L_i, \frac{h_m}{r}\rangle$ (i) $\langle e^{-r/2}L_i, e^{-r/2}L_j\rangle$.

Obviously (a) and (i) are trivial by (6.10). The product (e) also follows easily using the recurrence relation (6.12) to write L_n^2 and L_m^2 as sums of L_j's and then using (6.10). By symmetry, the only remaining cases are (b), (c) and (f). For (b) first note that the recurrence relation (6.11) gives

$$L_n^2 L_m^2 = \left(\sum_{i=0}^{n} L_i^1\right)\left(\sum_{j=0}^{m} L_j^1\right).$$

Combining this with the orthogonality of the L_i^1 with respect to the weight function re^{-r} we can drop the cross terms in the product $L_n^2 L_m^2$ to get

$$\begin{aligned}\left\langle h_n, \frac{h_m}{r}\right\rangle &= \int_0^\infty re^{-r} L_n^2(r) L_m^2(r)\,\mathrm{d}r \\ &= \int_0^\infty re^{-r} \sum_{i=0}^{\min(n,m)} (L_i^1)^2\,\mathrm{d}r \\ &= \sum_{i=0}^{\min(n,m)} \frac{\Gamma(i+2)}{i!} \\ &= \frac{1}{2}(k+1)(k+2),\end{aligned}$$

where $k \coloneqq \min(n,m)$. For (c), first note that by using (6.13) and then (6.11) we have

$$rL_n^2 L_j = (n+1)(L_n^1 - L_{n+1}^1)L_j = -(n+1)L_{n+1}L_j,$$

and so

$$\begin{aligned}\left\langle h_n, e^{-r/2} L_j\right\rangle &= \int_0^\infty re^{-r} L_n^2(r) L_j(r)\,\mathrm{d}r \\ &= -(n+1)\int_0^\infty e^{-r} L_{n+1} L_j\,\mathrm{d}r \\ &= -(n+1)\cdot\delta_{j,n+1},\end{aligned}$$

the last equality following from (6.10). Finally for (f), using (6.20) we have

$$\begin{aligned}\left\langle \frac{h_n}{r}, e^{-r/2} L_j\right\rangle &= \int_0^\infty e^{-r} L_n^2(r) L_j(r)\,\mathrm{d}r \\ &= \sum_{i=0}^{n} (n-(i-1))\delta_{ij} \\ &= (n-(j-1))\cdot I(j\le n),\end{aligned}$$

where $I(j \le n) = 1$ if $j \le n$ and $I(j \le n) = 0$ otherwise. Thus we have shown that all of the inner products (a)-(i), and hence (6.25), can be computed exactly from the indices n, m, i, j by using the recurrence relations for the generalised Laguerre polynomials. It follows the preceding discussion that all inner products on one of the three forms in (6.23) can be computed to arbitrary precision given $n, m, \lambda_n, \lambda_m$, assuming the square roots that appear in the constants N_n can be computed to arbitrary precision. As we remarked using the observation (6.22), this implies that the inner products (6.21) can be computed to arbitrary precision for any ψ_n, ψ_m. In conclusion, we have proved:

Proposition 6.7. *Let $H \in \Omega$ as in (6.7) and let $\psi_n, \psi_m \in \mathcal{B}_{\mathcal{H}'}$. Then assuming square roots of positive integers can be computed to an arbitrary level of precision using finitely many arithmetic operations and comparisons, then the inner product*

$$\langle (H - zI)\psi_m, (H - zI)\psi_n \rangle$$

can be computed to any level of precision using finite arithmetic means.

6.3 Main result on Dirac operators with infinite mass boundary conditions

With all the necessary ingredients in place we can string everything together and give the main result for Dirac-Coulomb operators with infinite mass boundary conditions. For completeness, we discuss how Algorithm 1 can be implemented in this specific case in Appendix A.2 and provide pseudocode in Algorithm 2.

Theorem 6.8. *Define the computational problem $\{\Xi, \Omega, \Lambda, \mathcal{M}\}$ as follows.*

(i) *The domain Ω is as in (6.7).*

(ii) *The evaluation functions Λ are the matrix evaluations $f_{i,j} \colon H \mapsto \langle H\psi_j, \psi_i \rangle$ and $g_{i,j} \colon H \mapsto \langle H\psi_j, H\psi_i \rangle$ with $\psi_i, \psi_j \in \mathcal{B}_{\mathcal{H}'}$.*

(iii) *The problem function is $\Xi(H) = \mathrm{Sp}(H)$.*

(iv) *$\mathcal{M}$ is the collection of closed subsets of $\mathbb{C}$ with the Attouch-Wets metric.*

Then $\{\Xi, \Omega, \Lambda, \mathcal{M}\} \in \Sigma_1$.

Proof. By assumption H is self-adjoint and in Lemma 6.6 we showed that the linear span of the basis $\mathcal{B}_{\mathcal{H}'}$ is a core for H. By Lemma 3.25, the sequence of functions $\Psi_n(z, H) = \sigma_1((H - zI)P_n)$ converges uniformly to $\gamma(z, H) \coloneqq \sigma_1(H - zI) = \|R(z, H)\|^{-1}$. Using Proposition 6.7 the inner products $\langle (H - zI)\psi_m, (H - zI)\psi_n \rangle$ can be computed up to arbitrary precision with finitely many arithmetic operations and comparisons. Then by Lemma 5.3, given $z \in \mathbb{C}$, for each $n \in \mathbb{N}$ we can compute $v_n(z, H)$ using finitely many arithmetic operations and comparisons such that

$$|v_n(z, H)^2 - \Psi_n(z, H)^2| \leq \frac{1}{n^2}. \tag{6.26}$$

Then $|v_n(z, H) - \Psi_n(z, H)| \leq 1/n$ and since $\Psi_n(z, H) \to \gamma(z, H)$ locally uniformly, so will $v_n(z, H)$ eventually. The convergence will also be monotone. In order to ensure that the convergence is *from above*, let

$$\gamma_n(z, H) \coloneqq v_n(z, H) + 1/n.$$

Then the sequence of functions $\gamma_n(z, H)$ satisfies the conditions in Proposition 3.29. Hence the sets $\Gamma_n(H)$ defined in Algorithm 1 will converge to $\mathrm{Sp}(H)$ in the Attouch-Wets metric with error control as described by the Σ_1-class in Definition 3.18. ■

CHAPTER 7

Concluding Remarks

In this thesis we have addressed the problem of computing the spectrum of a linear operator acting on an infinite dimensional Hilbert space. In Chapter 3 we gave sufficient conditions under which there exists a sequence of algorithms (a Σ_1-tower) which estimates the spectrum of a given operator with rigorous error control, using only arithmetic operations and comparisons. In Chapter 5 and Chapter 6 we showed how these conditions can be met for Dirac operators in three and two dimensions with different potentials. In particular, we proposed a numerical algorithm which can be used to find any points residing in the discrete spectrum $\mathrm{Sp}_{\mathrm{disc}}(H) \subset (-m, m)$ of the Dirac operator with infinite mass boundary conditions studied in [7].

Due to limitations in time, numerical implementation of the constructed algorithms will be a topic of further work. In particular, we note [7, Theorem 1.11] which says that in the case described by Theorem 6.3(ii) for any $\lambda \in (-m, m)$, there is a self-adjoint extension of $H_{\min}$ with λ as an eigenvalue. The Σ_1-algorithm in Section 6.3 provides a way to determine numerically how the complementary problem of how the spectrum behaves inside $(-m, m)$ in the case of Theorem 6.3(i), which is left unresolved by [7]. Since the search space $(-m, m)$ is bounded, the algorithm will eventually find every point within the spectrum. A natural extension of this thesis would be the application of the techniques used in Chapter 5 and Chapter 6 to other classes of Dirac operators, see e.g [31, Chapter 4]. One pertinent application is Dirac operators om domains in $\mathbb{R}^3$ which may have a gap in the essential spectrum [4, 22].

The presented results are in a way quite surprising. As remarked in Chapter 3, even for bounded diagonal operators we cannot achieve anything better than a Σ_1-classification, if the algorithm can only access a finite amount of matrix elements at a time. For compact operators, even if self-adjoint, a Σ_1-result is not possible without more information, and two limits are needed in order to achieve any error control. Thus, the problem of computing the spectrum for the Dirac operators in Chapter 5 given point samples from the potential, or the Dirac operator with infinite mass boundary conditions in Chapter 6 with access to matrix elements with respect to a cleverly chosen basis, is in some sense no more difficult than computing the spectrum of a bounded diagonal operator.

APPENDIX A

Arithmetic Algorithms

A.1 Linear algebra: Computing singular values

For a matrix $B \in \mathbb{C}^{n\times n}$ we denote its smallest singular value by $\sigma_1(B)$. This is defined as the square root of the smallest eigenvalue of the matrix B^*B, which is self-adjoint and positive semi-definite since

$$\min_{x\in\mathbb{C}^n} \langle B^*Bx, x\rangle = \min_{x\in C^n} \|Bx\|^2 \geq 0.$$

Next we give some simple results that are necessary to build the Σ_1-algorithms, adapted from [5].

Proposition A.1. *Given a matrix $B \in \mathbb{C}^{n\times n}$ with rational entries and any rational number $\eta > 0$, one can determine from the entries of B whether or not $\sigma_1(B) > \eta$ using finitely many arithmetic operations and comparisons.*

Proof. By definition, $\sigma_1(B) > \eta$ if and only if the smallest eigenvalue of B^*B is greater than η^2. This is the case if and only if the matrix $C \coloneqq B^*B - \eta^2 I$ is positive definite. To see this, observe that

$$\begin{aligned} &\langle (B^*B - \eta^2 I)x, x\rangle > 0 \text{ for all } x \neq 0 \\ \iff &\|Bx\|^2 > \eta^2\|x\|^2 \text{ for all } x \neq 0. \end{aligned}$$

Now the eigenvalues of B^*B are exactly the numbers $\bar{\lambda}\lambda = |\lambda|^2$ where λ is an eigenvalue of B. And clearly, $\|Bx\|^2 > \eta^2\|x\|^2$ for all $x \neq 0$ if and only if $|\lambda|^2 > \eta^2$ for all eigenvalues λ of B, showing the claimed equivalence. From linear algebra (see e.g. [28, p. 353]) we know that C is positive definite if and only if all of the pivot elements after row reduction (without row exchanges) are strictly positive. Row reduction of C is certainly done with finitely many arithmetic operations on the entries of B, during which the appearance of a pivot less than or equal to zero means C is not positive definite. If all pivots are greater than zero after row reduction, then C is positive definite. ■

With the above result established, it is not surprising that one can compute the smallest singular value of a matrix using finitely many arithmetic operations and comparisons. Before giving the formal proof, we recall that for $A, B \in \mathbb{C}^{m\times n}$, $|\sigma_1(A) - \sigma_1(B)| \leq \|A - B\|$. Also recall that the operator norm is bounded above by the Frobenius norm, i.e., $\|A\| \leq \sqrt{\sum_{ij}|A_{ij}|^2}$.

Proposition A.2. *Let $B \in \mathbb{C}^{n\times n}$ be a matrix such that we can compute each of its entries to arbitrary precision using finitely many arithmetic operations and comparisons. Then for any $\varepsilon > 0$ we can compute $\sigma_1(B)$ to within an accuracy of ε using finite arithmetic means.*

Proof. We can assume without loss of generality that $\varepsilon \in \mathbb{Q}$. By assumption we can compute an approximation $\widehat{B}$ of B with rational entries such that each entry has error less than $\varepsilon/(2n)$ so that $\|B - \widehat{B}\| < \varepsilon/2$. It follows that $|\sigma_1(B) - \sigma_1(\widehat{B})| < \varepsilon/2$. If we can compute the singular value $\sigma_1(\widehat{B})$ to within $\varepsilon/2$, then the proposition follows. Now compute $M \in \mathbb{N}$ such that $M^{-1} < \varepsilon/2$. Then by Proposition A.1 we can iteratively compute via finite arithmetic means the smallest number $k \in \mathbb{N}$ such that $\sigma_1(\widehat{B}) \leq k/M$. Then we can use k/M as our estimate for $\sigma_1(\widehat{B})$ since

$$\frac{k-1}{M} < \sigma_1(\widehat{B}) \leq \frac{k}{M}.$$

Hence k/M is at most $1/M$ greater than $\sigma_1(\widehat{B})$ and since $1/M < \varepsilon/2$, the result follows. ∎

Of course there are far more efficient procedures for computing eigenvalues and singular values. The purpose here is simply to show that all the algorithms constructed can be rigorously performed by an arithmetic algorithm, i.e., a Turing machine.

A.2 Dirac operator with infinite mass boundary conditions

In this section, we formulate precisely the Σ_1-algorithm in Proposition 3.29 as it applies to the Dirac operator H in Theorem 6.8. We neglect the constants $m = \nu = 1$ in going forward as they can easily be incorporated into the algorithms with any values. In Definition 3.26 we have $g(x) = x$ and so $h_n\colon [0,\infty) \to [0,\infty)$ in (3.12) is simply given by

$$h_n(y) = \min_{k\in\mathbb{N}}\{k/n\colon k/n > y\}, \tag{A.1}$$

and can easily be computed using finitely many arithmetic operations and comparisons for any given $y \in \mathbb{Q}$. In the computation of $\Gamma_n(H)$, given z we need to compute the estimate $v_n(z,H)$ of $\Psi_n(z,H) = \sigma_1((H - zI)P_n)$, which we do using Lemma 5.3. Specifically, we consider the self-adjoint matrix $\widetilde{W}_n$ with entries $\{\widetilde{W}_n(z)\}_{ij} = \langle (H - zI)\psi_i, (H - zI)\psi_j\rangle$ for $1 \leq i,j \leq n$, where $\psi_i, \psi_j \in \mathcal{B}_{\mathcal{H}'}$. As shown in Section 6.2 $\{\widetilde{W}_n(z)\}_{ij}$ can be computed to any precision using the recurrence relations of the Laguerre polynomials, assuming square roots of positive integers can be computed to any precision. Therefore, suppose we have a subroutine InProd(n) which given any $n \in \mathbb{N}$ returns a rational matrix W_n (which we can assume is Hermitian), approximating $\widetilde{W}_n$ with entrywise error $E^n_{i,j}(z) \leq 1/(8n^3)$. Then $\left|\sigma_1(\widetilde{W}_n) - \sigma_1(W_n)\right| \leq \sqrt{\sum_{ij}(E^n_{i,j})^2} \leq 1/(8n^2)$. Note that, $\sigma_1(W_n) > k/(8n^2)$ if and only if $W_n^2 - k^2/(8n^2)^2 I$ is positive definite. An the previous section, we can find the smallest k such that $\sigma_1(W_n) < k/(8n^2)$. Then, setting $v_n(z,H) = \sqrt{k/(8n^2)}$, we have $\left|v_n(z,H)^2 - \sigma_1(W_n)\right| \leq 1/(8n^2)$, which in turn

implies $|v_n(z,H)^2 - \Psi_n(z,H)^2| \leq 1/(4n^2)$ and so $|v_n(z,H) - \Psi_n(z,H)| \leq 1/(2n)$. Using e.g. a power series expansion of the square root function on $\mathbb{R}_+$ with a Lagrange error bound, we can finally compute an estimate $\widehat{v}_n(z,H)$ of $v_n(z,H)$ such that $|\widehat{v}_n(z,H) - v_n(z,H)| \leq 1/(2n)$ and so $|\widehat{v}_n(z,H) - \Psi_n(z,H)| \leq 1/n$. Pseudocode for this procedure and for the estimation of the spectrum is given in Algorithm 2.

Algorithm 2: The routine `CompSpec(`n`)` computes the estimates $\Gamma_n(H) \to \mathrm{Sp}(H)$, also providing an error bound E_n. Note that the grid is contained in the real interval $(-1,1)$. The subroutine `IsPosDef()` checks whether a matrix is positive definite and can be implemented in many different ways. `InProd(`n`)` computes the matrix W_n estimating $\widetilde{W}_n$ with entrywise error less than $1/(8n^3)$. Any desired values of m and ν in the definition of H can be incorporated here. The subroutine `sqrt(`q`,` ε`)` computes the square root of the positive rational number q with error less than ε. Note that in practice, the while loop in `DistSpec(`n`)` would be replaced by a faster binary search method.

Function `DistSpec(`n, z`)`**:**
- **Input:** $n \in n, z \in \mathbb{R}$
- **Output:** $v \in \mathbb{R}_+$, approximation to $\Psi_n(z,H)$
- $W_n =$ `InProd(`n`)`
- $q = 1, k = 0$
- **while** $q = 1$ **do**
  - $k = k + 1$
  - $q =$ `IsPosDef(`$W_n^2 - \frac{k^2}{(8n^2)^2} I$`)`
- **end**
- $v =$ `sqrt(`$k/(8n^2)$`,` $1/(2n)$`)`

return

Function `CompSpec(`n`)`**:**
- **Input:** $n \in n$.
- **Output:** Approximation $\Gamma_n \subset \mathbb{C}$, error estimate $E_n \in \mathbb{R}_+$.
- $G = \frac{1}{n}\mathbb{Z} \cap B_1(0)$
- **for** $z \in G$ **do**
  - $\gamma_n =$ `DistSpec(`n`,` z`)` $+ 1/n$
  - **if** $\gamma_n(z) \leq (|z|^2 + 1)^{-1}$ **then**
    - $\Upsilon \coloneqq B_{h_n(\gamma_n(z))}(z) \cap G$
    - **for** $w \in \Upsilon$ **do**
      - $F_w =$ `DistSpec(`w`)` $+ 1/n$
    - **end**
    - $M_z = \{w \in \Upsilon \colon F_w = \min_{v \in \Upsilon} F_v\}$
  - **else**
    - $M_z = \emptyset$
  - **end**
- **end**
- $\Gamma_n \coloneqq \cup_{z \in G} M_z$
- $E_n = \max_{z \in \Gamma_n} h_n(\gamma_n(z))$

return

List of Symbols and Notation

$\mathcal{H}$	separable Hilbert space
$\mathcal{L}(\mathcal{H})$	set of linear operators on $\mathcal{H}$
$\mathcal{B}(\mathcal{H})$	set of bounded operators on $\mathcal{H}$
$\mathcal{C}(\mathcal{H})$	set of closed, densely defined operators on $\mathcal{H}$
$\mathbb{C}^{m\times n}$	set of $m \times n$ complex matrices
$\mathbb{R}_+$	positive real numbers
$B_r(x)$	closed ball (in a metric space) of radius $r > 0$ centered at x
$\mathrm{cl}(S)$	closure of a set S in a topological space
$d_H(B, C)$	Hausdorff distance between non-empty compact sets B, C
$d_{\mathrm{AW}}(B, C)$	Attouch-Wets distance between non-empty closed sets B, C
$B + C$	set of all $b+c$ where $b \in B$ and $c \in C$ for non-empty $B, C \subset \mathbb{C}$
$\mathcal{D}(A)$	domain of operator A
$\mathcal{R}(A)$	range of operator A
$\mathcal{N}(A)$	kernel of operator A
$\mathcal{G}(A)$	graph of operator A
A^*	the adjoint of a operator A
$\overline{A}$	closure of operator A
$A \restriction D$	restriction of operator A to subspace $D \subset \mathcal{D}(A)$
$\rho(A)$	resolvent set of operator A
$\mathrm{Sp}(A)$	spectrum of operator A, $\mathrm{Sp}(A) = \mathbb{C} \setminus \rho(A)$
$\mathrm{Sp}_{\mathrm{ess}}(A)$	essential spectrum of operator A
$\mathrm{Sp}_{\mathrm{disc}}(A)$	discrete spectrum of operator A

$\sigma_1(M)$	smallest singular value of matrix $M \in \mathbb{C}^{m\times n}$, generalised to operators in Equation (3.9)
$\mathcal{H} \otimes \mathcal{H}'$	tensor product of Hilbert spaces $\mathcal{H}$ and $\mathcal{H}'$
$\mu \otimes \nu$	product measure generated by measures μ and ν
$\operatorname{dist}(z, C)$	distance from point z to closed set $C \subset \mathbb{C}$
$L^2(\Omega; \mathbb{C}^k)$	Hilbert space of functions with k components, each an element of $L^2(\Omega, \mathrm{d}\mu) = L^2(\Omega; \mathbb{C})$ for a locally compact Hausdorff space Ω
$H^1(\Omega; \mathbb{C}^k)$	set of functions with k components, each an element of first Sobolev space, $H^1(\Omega) \subset L^2(\Omega; \mathbb{C})$
$C_c^\infty(\Omega; \mathbb{C}^k)$	set of functions with k components, each smooth and compactly supported on some domain $\Omega \subset \mathbb{R}^d$
$\mathcal{S}(\Omega; \mathbb{C}^k)$	set of functions with k components, each a component of the Schwartz space $\mathcal{S}(\Omega; \mathbb{C})$ for some domain $\Omega \subset \mathbb{R}^d$
$L^2\big(\mathbb{R}, w(x)\,\mathrm{d}x\big)$	the weighted L^2-space with measure $w(x)\,\mathrm{d}x$
δ_{ij}	Kronecker delta symbol

Bibliography

[1] Aberth, O. 'The Failure In Computable Analysis of a Classical Existence Theroem for Differential Equations'. In: *Proceedings of the American Mathematical Society* vol. 30, no. 1 (1971), p. 6.

[2] Arveson, W. 'The Role of C^*-Algebras in Infinite-Dimensional Numerical Linear Algebra.' In: *C^*-algebras: 1943–1993 (San Antonio, TX, 1993)* vol. volume 167 of Contemp. Math. (1994), pp. 114–129.

[3] Beer, G. and Diconcilio, A. 'Uniform Continuity on Bounded Sets and the Attouch-Wets Topology'. In: *Proceedings of the American Mathematical Society* vol. 112, no. 1 (1991), pp. 235–243.

[4] Behrndt, J., Holzmann, M. and Mas, A. 'Self-Adjoint Dirac Operators on Domains in $\mathbb{R}^3$'. In: *Annales Henri Poincare* vol. 21, no. 8 (2020), pp. 2681–2735.

[5] Ben-Artzi, J. et al. *Computing Spectra – On the Solvability Complexity Index Hierarchy and Towers of Algorithms*. June 2020. arXiv: *1508.03280*.

[6] Blümlinger, M. and Tichy, R. F. 'Topological Algebras of Functions of Bounded Variation I'. In: *manuscripta mathematica* vol. 65, no. 2 (June 1989), pp. 245–255.

[7] Cassano, B., Gallone, M. and Pizzichillo, F. *Dirac-Coulomb Operators with Infinite Mass Boundary Conditions in Sectors*. Feb. 2022. arXiv: *2202.13180*.

[8] Colbrook, M. J. and Hansen, A. C. *The Foundations of Spectral Computations via the Solvability Complexity Index Hierarchy: Part I*. Aug. 2020. arXiv: *1908.09592*.

[9] Davies, E. B. *Linear Operators and Their Spectra*. Cambridge Studies in Advanced Mathematics 106. Cambridge: Cambridge University Press, 2007.

[10] Dirac, P. A. M. *The Principles of Quantum Mechanics*. 4. ed. (rev.), repr. International Series of Monographs on Physics 27. Oxford: Clarendon Press, Oxford University Press, 2010.

[11] Dong, S.-H. 'Realization of the Dynamical Group for the Generalized Laguerre Functions'. In: *Computers & Mathematics with Applications* vol. 47, no. 6-7 (Mar. 2004), pp. 1035–1039.

[12] Evans, L. C. *Partial Differential Equations*. 2nd ed. Graduate Studies in Mathematics v. 19. Providence, R.I: American Mathematical Society, 2010.

[13] Folland, G. B. *Fourier Analysis and Its Applications*. The Wadsworth & Brooks/Cole Mathematics Series. Pacific Grove, Calif: Wadsworth & Brooks/Cole Advanced Books & Software, 1992.

[14] Folland, G. B. *Real Analysis: Modern Techniques and Their Applications*. 2nd ed. Pure and Applied Mathematics. New York: Wiley, 1999.

[15] Grubb, G. *Distributions and Operators*. Graduate Texts in Mathematics 252. New York: Springer, 2009.

[16] Hansen, A. C. 'On the Solvability Complexity Index, the n-Pseudospectrum and Approximations of Spectra of Operators'. In: *Journal of the American Mathematical Society* vol. 24, no. 1 (Jan. 2011), p. 44.

[17] Indritz, J. 'An Inequality for Hermite Polynomials'. In: *Proceedings of the American Mathematical Society* vol. 12, no. 6 (1961), pp. 981–983.

[18] Katō, T. *Perturbation Theory for Linear Operators*. Classics in Mathematics. Berlin: Springer, 1995.

[19] Le Treust, L. and Ourmières-Bonafos, T. 'Self-Adjointness of Dirac Operators with Infinite Mass Boundary Conditions in Sectors'. In: *Annales Henri Poincaré* vol. 19, no. 5 (May 2018), pp. 1465–1487.

[20] Leary, C. C. and Kristiansen, L. *A Friendly Introduction to Mathematical Logic*. Second edition. New York: SUNY Geneseo at Geneseo, 2015.

[21] Lindstrøm, T. *Spaces: An Introduction to Real Analysis*. Pure and Applied Undergraduate Texts volume 29. Providence, R.I: American Mathematical Society, 2017.

[22] Mas, A. and Pizzichillo, F. 'The Relativistic Spherical δ -Shell Interaction in R 3 : Spectrum and Approximation'. In: *Journal of Mathematical Physics* vol. 58, no. 8 (Aug. 2017), p. 082102.

[23] Niederreiter, H. *Random Number Generation and Quasi-Monte Carlo Methods*. CBMS-NSF Regional Conference Series in Applied Mathematics 63. Philadelphia, PA: Society for Industrial and Applied Mathematics, 1992.

[24] Reed, M. and Simon, B. *Methods of Modern Mathematical Physics I, Functional Analysis*. New York: Academic Press, 1972.

[25] Reed, M. and Simon, B. *Methods of Modern Mathematical Physics II, Fourier Analysis, Self-Adjointness*. San Diego: Academic press, 1975.

[26] Schmüdgen, K. *Unbounded Self-Adjoint Operators on Hilbert Space*. Graduate Texts in Mathematics 265. Dordrecht New York: Springer, 2012.

[27] Soare, R. *Turing Computability: Theory and Applications*. New York, NY: Springer Berlin Heidelberg, 2016.

[28] Strang, G. *Linear Algebra and Its Applications*. 4th ed. Belmont, CA: Thomson, Brooks/Cole, 2006.

[29] Szegő, G. *Orthogonal Polynomials*. 4th ed. Colloquium Publications - American Mathematical Society v. 23. Providence, R.I: American Mathematical Society, 1939.

[30] Takhtadzhian, L. A. *Quantum Mechanics for Mathematicians*. Graduate Studies in Mathematics v. 95. Providence, R.I: American Mathematical Society, 2008.

[31] Thaller, B. *The Dirac Equation*. Texts and Monographs in Physics. Berlin ; New York: Springer-Verlag, 1992.

[32] Turing, A. M. 'On Computable Numbers, with an Application to the Entscheidungsproblem'. In: *Proceedings of the London Mathematical Society* vol. s2-42, no. 1 (1937), pp. 230–265.