SISSA

Scuola
Internazionale
Superiore di
Studi Avanzati

Mathematics Area – Ph.D. Course in

Mathematical Analysis, Modelling, and Applications

On Krylov Methods in Infinite-dimensional Hilbert Space

Candidate:
Noè Angelo Caruso

Supervisors:
Alessandro Michelangeli
Paolo Novati

Academic Year 2018-2019

A mia sorella Dalila

Abstract

This thesis contains the development of key features for the solution to inverse linear problems $Af = g$ on infinite-dimensional Hilbert space $\mathcal{H}$ using projection methods. Particular attention is paid to *Krylov subspace methods*. Intrinsic, key operator-theoretic constructs that guarantee the 'Krylov solvability' of the problem $Af = g$ are developed and investigated for this class of projection methods. This theory is supported by numerous examples, counterexamples, and some numerical tests. Results for both bounded and unbounded operators on general Hilbert spaces are considered, with special attention paid to the Krylov method of conjugate-gradients in the unbounded setting.

Acknowledgements

I am most grateful to my supervisor Prof. Alessandro Michelangeli who always encouraged me to keep going, who was always inquisitive and insightful, and was such an invaluable source of guidance and inspiration for me.

Also deserving of special thanks and gratitude are my two dear friends Ornela Mulita and Zakia Zainib who supported me throughout my studies and whose encouragement motivated me throughout my most difficult times.

Declaration

This thesis is submitted in partial fulfilment of the degree Doctor of Philosophy in Mathematical Analysis, Modelling, and Applications. The work presented in this thesis is, except where acknowledged in the customary manner, to the best of my knowledge, original and has not been submitted in whole or in part for a degree in any university.

Noè Angelo Caruso
August, 2019

Contents

List of Figures

Commonly used symbols

Symbol	Description
$\mathcal{H}$	Abstract Hilbert space
$\langle\cdot,\,\cdot\rangle$	Scalar product on $\mathcal{H}$, antilinear in the first argument
$\lVert\cdot\rVert_{\mathcal{H}}$	Metric for the Hilbert space $\mathcal{H}$, $\lVert\cdot\rVert_{\mathcal{H}}^2 = \langle\cdot,\,\cdot\rangle$.
$\lVert\cdot\rVert_{\text{op}}$	Operator norm
$\mathscr{B}(\mathcal{H})$	Collection of bounded operators on $\mathcal{H}$
$\mathscr{C}(\mathcal{H})$	Collection of closed operators on $\mathcal{H}$
A	Linear, closed operator on $\mathcal{H}$
A^*	Adjoint of the operator A
$\mathbb{1}$	Identity on $\mathcal{H}$
$\mathbb{O}$	Zero operator on $\mathcal{H}$
$\rho(A)$	Resolvent set for A
$\sigma(A)$	Spectrum of A
$\mathcal{R}\,(A,\,\zeta)$	Resolvent operator $(\zeta\mathbb{1} - A)^{-1}$ for $\zeta \in \rho(A)$
$\mathcal{D}\,(A)$	Domain of A
$\text{ran}A$	Range of A
$G\,(A)$	Graph space of A
$\lVert\cdot\rVert_{G(A)}$	Graph norm of A
$C(X,Y)$	Continuous functions between topological spaces X and Y
$C_c(X,Y)$	Continuous functions between topological spaces X and Y with compact support on X
$C_0(X,Y)$	Continuous functions between topological spaces X and Y that vanish at infinity
$C^\infty(X,Y)$	Smooth functions between spaces $X \subset \mathbb{R}$ or $\mathbb{C}$, and $Y \subset \mathbb{R}$ or $\mathbb{C}$
$\mathcal{C}^\infty\,(A)$	Space of vectors $g \in \mathcal{H}$ such that $g \in \bigcap_{n\in\mathbb{N}_0} \mathcal{D}\,(A^n)$
$\mathbb{C}$	Complex numbers
$\mathbb{N}, \mathbb{N}_0$	Natural numbers, and positive integers respectively
$\mathbb{Z}$	Integers
$\mathbb{R}$	Real numbers
$\mathbf{E}\,(\cdot)$	Spectral measure
$\lvert\psi\rangle\,\langle\psi\rvert$	Rank-1 orthogonal projection onto $\psi \in \mathcal{H}$

Chapter 1

Introduction and Scope

This thesis investigates the mathematical framework, key features, the discretisation setting, and the approximability of linear inverse problems in *infinite*-dimensional Hilbert space. Special attention is paid to applications of the renowned Krylov subspace methods, and in particular the issue of 'Krylov solvability' of a given inverse problem.

Several convergence criteria are stated and investigated for general projection methods for linear inverse problems in the abstract operator-theoretic setting, with particular attention devoted to the class of Krylov subspace methods. These results combine together to give new theoretical insights that are useful to study various classes of inverse problems, especially in the unbounded operator setting, and give confidence that Krylov subspace methods can be sound numerical techniques for solving said problems.

Krylov subspace methods are ubiquitous in scientific computing and have been described as one of '...the 10 Algorithms with the greatest influence on the development and practice of science and engineering in the 20th century.' [22]. These methods are so popular that they often appear in monographs dedicated to other numerical topics, such as finite element methods [74, 75], that are themselves particularly useful in modelling physical systems (e.g. [18, 60]).

Still at an informal level, the overarching objective of this work will now be discussed, with the more structured outline and major contributions in the following sections.

The first core notion to be discussed within this thesis is the linear inverse problem. In abstract Hilbert spaces, the preferred setting in this work, the linear inverse problem is the problem of finding solution(s) f to

$$Af = g\,, \tag{1.1}$$

where A is a closed, densely defined, linear operator on a Hilbert space $\mathcal{H}$ equipped with norm $\|\cdot\|_{\mathcal{H}}$ and scalar product $\langle\cdot,\,\cdot\rangle$, and $g \in \mathcal{H}$ is a vector.

The second core notion involves the numerical strategy of Krylov subspace methods to find solution(s) $f \in \mathcal{H}$ to (1.1). Throughout this thesis, $\langle\cdot,\,\cdot\rangle$ is taken as antilinear in the first argument and linear in the second. In contrast to much of the established literature on Krylov methods, here the operator A is, without loss of generality, assumed to be a *genuinely* infinite-dimensional operator. That is, as is conventional in [88, Section 1.4], A is *not* reduced to $A = A_1 \oplus A_2$ by an orthogonal direct sum decomposition $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$ where $\dim \mathcal{H}_1 < \infty$, $\dim \mathcal{H}_2 = \infty$ and $A_2 = \mathbb{O}$.

If there exists a solution $f \in \mathcal{H}$ to (1.1), i.e., $g \in \mathrm{ran}A$, the problem is called *solvable*. In addition, if the solution f is unique, i.e., A is injective, the problem is called *well-defined* and one refers to f as the *exact* solution to (1.1). Finally, if (1.1) is well-defined and the exact solution f depends continuously on the given datum g, the problem is called *well-posed*.

There are several types of different Krylov subspace methods, but *standard* Krylov (or simply *Krylov*) approximation methods search for the solution(s) to (1.1) in the *cyclic space* generated by the operator A and the vector g, i.e., the subspace generated by the linear span of the vectors $g, Ag, A^2g, \dots$. In fact, Krylov methods may be seen in the setting of *iterative* methods, or general *projection* methods. Framing Krylov methods in both these forms will be done in this thesis, using the advantages of each formulation as appropriate in the analysis.

In the setting of solving (1.1), applications and convergence properties of Krylov methods are now the subject of a well-established, classical area of literature and are treated in several well-known monographs [87, 55, 25, 41, 53]. Most of the current literature considers the linear inverse problem formulated in finite-dimensions, i.e., when $\dim \mathcal{H} < \infty$. There is a smaller, yet still significant, part of the literature on the infinite-dimensional setting, in particular [21, 49, 64], and more recently [44, 69, 35].

In this spirit, it is worth explicitly mentioning the work of Nemirovskiy and Polyak [64, 65] for conjugate-gradient methods as applied to bounded, self-adjoint, positive operators $A \geq \mathbb{O}$. This work [64] definitively established the strong convergence of the sequence of numerical approximants generated by the algorithm at each step, i.e., $(f^{[N]})_{N \in \mathbb{N}} \subset \mathcal{H}$, to a single solution of the solvable problem (1.1). That is, $\|f^{[N]} - f\|_{\mathcal{H}} \to 0$ as $N \to \infty$ for some $f \in \mathcal{H}$ a solution to the solvable problem (1.1). Moreover, the rate of convergence presented in [64] was proven in a follow-up work [65] to be the *optimal* rate for the entire class of operators considered therein.

Although there are studies of Krylov methods in the infinite-dimensional setting, currently a *systematic* study of these general methods is lacking. Moreover, many of these infinite-dimensional studies require further assumptions on the underlying operator such as self-adjointness or compactness, and only examine *specific* Krylov methods.

Currently, there is a lack of the classification of the general operator-theoretic mechanisms that ensure whether the treatment of (1.1) is appropriate using these methods. An appropriate treatment of the solvable problem (1.1) would require that there *is* in fact a solution $f \in \mathcal{H}$ that may be arbitrarily well approximated by linear combinations of vectors in the Krylov subspace. Under this setting, any solution to (1.1) $f \in \mathcal{H}$ with this property is referred to as being a *Krylov solution* to problem (1.1); and more generally (1.1) is called *Krylov solvable*. Informally, the *Krylov solvability* or *lack of Krylov solvability* of (1.1) occurs when a solution exists in the closure of the associated Krylov space or no such occurrence exists, respectively.

Most certainly, the issue of Krylov solvability is non-trivial for the solution to a solvable linear system (1.1). For example, one may consider the case where A is a bounded, injective operator with non-dense range, and the solution f is perpendicular to $\operatorname{ran} A$. Obviously the linear span of the vectors $g, Ag, A^2g, \ldots$ cannot approximate f within an arbitrary tolerance, as the

Krylov space is contained in ran$A$. As such, still at an abstract level, there do exist well-defined problems (1.1) such that they are *not* solvable using standard Krylov methods.

In this respect, the primary aim of this thesis is to develop the appropriate notions of numerical convergence and the *suitability* of using projection methods. More specifically, operator-theoretic constructions for *necessary and sufficient* conditions that ensure Krylov solvability are developed along with convergence properties of general projection methods in their abstract formulation. The general projection methods presented here are a suitable generalisation outside the standard framework of Petrov-Galerkin methods, without underlying assumptions on the density of the projection bases, and further removing the assumption of the solvability of the truncated problem at the finite-dimensional level. In addition to this, the convergence properties for the class of Krylov methods known as conjugate-gradient style methods is generalised to the setting of unbounded operators.

Krylov solvability is a major theme that will be recurrent throughout this thesis, and is developed within the setting of standard, or *polynomial*, Krylov spaces; but some aspects of Krylov solvability are also touched on for *rational* Krylov spaces. The results stated herein are still suitable for use when $\dim \mathcal{H} < \infty$, or when the operator A *is* reduced with respect to a finite-dimensional subspace $\mathcal{H}_1 \subset \mathcal{H}$.

1.1 Thesis outline

This thesis begins with a brief background on Krylov methods in Chapter 2, along with their associated definitions and discussion of the relevant literature. The particular focus of the literature in this thesis is on Krylov methods in the context of the solvable linear inverse problem (1.1), within the realm of infinite-dimensional Hilbert spaces. In this setting, for historical and analytical reasons, the conjugate-gradient method holds a special place and is given extra attention. The focus in Chapter 2 remains on polynomial Krylov methods, and the discussions of the more recent rational Krylov methods is presented in later chapters, in context with theoretical results.

Following this review, Chapter 3 concerns development of the theory for the convergence of general projection methods, with attention also paid to *Galerkin* and *Petrov-Galerkin* methods. The underlying conditions and assumptions are explored that guarantee the *strong* or *weak* convergence of numerically approximated solutions of (1.1) to an actual solution. This theory is explored and developed with appropriate examples and counterexamples presented, along with some simple numerical tests.

The case where the vector g in the solvable problem (1.1) contains some extra 'noise' term is an area that shall only be lightly touched on in Chapter 3 within the context of projection methods. In fact, the process of attaining good estimates to the *true* solution in the presence noise is a well-established field known as *regularisation*. Some monographs that discuss this area in-depth include [25, 42, 93].

In Chapter 4 the operator-theoretic notions of Krylov solvability of the solvable problem (1.1) are developed for the class of *bounded* linear operators on $\mathcal{H}$, i.e., the algebra $\mathscr{B}(\mathcal{H})$. In particular, the operator-theoretic notion of the '*Krylov intersection*' is developed along with some interesting examples and counterexamples that unmask the theoretical constructions. This *Krylov intersection* construction is particularly important and is found to capture the *essence* of Krylov solvability at the most general level.

Chapter 5 contains the theory of Krylov solvability extended to the more general class of *densely defined, closed* linear operators on $\mathcal{H}$. In this setting, many of the theoretical results of the previous chapter are generalised, taking into account the unavoidable (and often subtle) domain issues that arise when considering unbounded operators. It is shown that the relevant operator theoretic constructions of the previous section are still valid, in a more generalised sense, further consolidating their nature as the *intrinsic* mechanisms of Krylov solvability.

To the current sensibilities of the literature, this setting of unbounded Krylov methods appears to be an area of little theoretical development for Krylov solvability at the infinite-dimensional level. Some initial steps have been made for linear differential operators [69, 35] particularly in the solution to highly oscillatory integrals [70, 71]. Within Chapter 5, some of the aspects of rational Krylov solvability are also discussed in context with the literature.

Following this study of abstract operator-theoretic mechanisms in the unbounded setting, Chapter 6 contains a concrete application to the conjugate-gradient method investigated for the entire class of *self-adjoint, positive* operators. In particular, a convergence result is obtained, showing that indeed the convergence of the numerical approximates to a solution is guaranteed under certain natural assumptions on the vector $g \in \mathrm{ran}A$. This, of course, has immediate applications to linear inverse problems arising from self-adjoint, positive differential operators.

Within Chapter 7, some future perspectives are lightly touched upon. Firstly, some future directions are given on the topic of Krylov solvability from an auxiliary perturbed inverse linear problem $A_\varepsilon f_\varepsilon = g$. In particular, it is planned to investigate suitable conditions under which in the limit of vanishing ε one may say that the Krylov solvability of $A_\varepsilon f_\varepsilon = g$ survives, where $A_{\varepsilon=0} = A$ for $A \in \mathscr{B}(\mathcal{H})$. Secondly, an application to unbounded linear inverse problems is planned for Friedrichs systems. Friedrichs systems are already treatable using finite element methods [27], however these methods require boundary conditions that reduce the problem to that of a *bounded* linear system with *everywhere defined bounded inverse* (i.e., coercivity). It is planned to investigate the Krylov solvability properties of these problems in the truly unbounded setting, also with the possibility of removing the coercivity assumption.

Finally, after Chapter 7, there are several appendices containing the most commonly used operators in this thesis along with their properties, some elements of operator and spectral theory, and functional analysis miscellanea.

1.2 Main results and contributions

The key results and contributions of this thesis are presented in Chapters 3 through 6. The major results and remarks from Chapter 3 are based on the work Caruso, Michelangeli, and Novati [17] and include the following.

- In general projection methods, there are always truncations that create unsolvable problems at the finite-dimensional level for every size of the truncation. These problems are mitigated by suitable assumptions on the method as well as the operator in question (e.g., approximability of the ambient Hilbert space, coercivity of the operator, etc).

- At the level of *compact* operators and under the assumption of the asymptotic consistency of the truncated problem, the phenomenon of the strong vanishing of the residual results in at least the component-wise vanishing of the error term. Weak vanishing of the error occurs under the further assumption of uniform boundedness of the numerical approximants.

- At the level of *general* bounded operators, the strong vanishing of the error and residual terms is dependent on the asymptotic solvability of the truncated problem coupled to the strong convergence of the numerical approximants. As compared to compact operators, now the control of the convergence is dependent on the stronger requirement of the convergence of the numerical approximants rather than the mere assumption of uniform boundedness.

The results and remarks from Chapter 4 are based on Caruso, Michelangeli, and Novati [16] and include the following points.

- In general, the solution to a well-defined linear inverse problem $Af = g$ may *not* be said to be in the closure of the Krylov space without further information. An explicit counter-example presented is that of the right-shift operator on $\ell^2(\mathbb{Z})$.

- If the Krylov subspace and its orthogonal complement are invariant under the action of the operator A (i.e., the Krylov space is *reduces* the operator A), then there exists a solution to the linear inverse problem in the closed Krylov subspace. Krylov reducibility always holds for bounded self-adjoint operators.

- In the general class of bounded normal operators, even if the linear inverse problem has Krylov solution, it does not guarantee Krylov reducibility. An explicit counter-example is provided using the multiplication operator $f \mapsto zf$ on $L^2(\Omega)$ for $\Omega \subset \mathbb{C}$ a suitable disc.

- The general operator-theoretic mechanism of Krylov solvability for an injective, bounded operator A, is that of the linear subspace known as *Krylov intersection*. The triviality of this subspace guarantees Krylov solvability for the linear inverse problem. The triviality of the Krylov intersection is also *equivalent* to the Krylov solvability under the condition that A is a bounded bijection.

- Under a lack of injectivity, should a Krylov solution exist and $\ker A \subset \ker A^*$, then it is guaranteed to be *unique*. Therefore, for all bounded, self-adjoint operators, the linear inverse problem for has a *unique* Krylov solution.

The results and remarks from Chapter 5 are the suitable generalisation of those from Chapter 4, and are based on the work [15]. The class of operators considered is that of the *closed, densely defined* operators in $\mathcal{H}$. The major findings include the following.

- Standard Krylov subspaces are ensured to be well-defined using suitably smooth vectors g, such that they remain in all powers of the operator A.

- Krylov reducibility still guarantees the Krylov solvability of the linear inverse problem, under an additional regularity assumption that the projection of the solution onto the closed Krylov subspace still remains within the domain of A.

- Similarly, under the projection condition on the solution described above, the triviality of the Krylov intersection still guarantees the existence of a Krylov solution for an injective operator. This shows that this mechanism still remains the intrinsic operator-theoretic mechanism of the Krylov solvability, and captures the *essence* of the Krylov solvability at the most general level.

- Under conditions of the lack of uniqueness of solutions, one still has that the uniqueness conclusions from Chapter 4 are the same.

- Owing to consequences of the possible unboundedness, one can no longer make such general statements about Krylov solvability of self-adjoint operators. Moreover, the Krylov reducibility may fail to hold for a general self-adjoint operator.

- *Rational* Krylov methods, built using general rational functions of the injective operator A, may exhibit Krylov solvability for the inverse problem from self-adjoint operators A. Explicit conditions are laid out guaranteeing the Krylov solvability that depend on the choice of the poles of the rational functions.

The specific study of the conjugate-gradients technique in Chapter 6 is presented in Caruso and Michelangeli [14] as applied to general *unbounded, self-adjoint, positive operators.* The main findings and contributions are summarised as follows.

- Under suitable assumptions on the regularity of the datum g, and the initial guess for the algorithm, the conjugate-gradient method is well-defined.

- The strong convergence of the numerical approximants to a single solution is guaranteed and shown in the proof of the main result, Theorem 6.4.1.

- The analysis is, as in the spirit of Nemirovskiy and Polyak [64], general enough to take into account not only *the* conjugate-gradient method, but *all* conjugate-gradient style methods that are formulated as the minimisation of an appropriate functional

  $$\rho_\theta(h) = \left\| A^{\theta/2}(h - \mathcal{P}_{\mathcal{S}}h) \right\|_{\mathcal{H}}^2 , \tag{1.2}$$

  where $\theta \geq 0$ and $\mathcal{P}_{\mathcal{S}}$ is the projection onto the manifold of solutions to (1.1), and h is taken in a space of suitably chosen vectors. Aside from the conjugate-gradient method ($\theta = 1$), other notable examples include: conjugate-gradients on the normal equations (CGNE); the least-square QR method (LSQR); minimal residual method (MINRES) applied to the class of unbounded, self-adjoint, positive operators; etc.

- The proof provided uses results from orthogonal polynomial theory in a novel way that show the convergence of numerical approximants in a more general setting that is not as restrictive as the original result proved by Nemirovskiy and Polyak [64].
- Required assumptions to ensure convergence in other norms, for example the energy (semi-)norm $\langle \cdot, \, A\cdot \rangle$ are considered and discussed.

Chapter 2

Background and Review

2.1 Introduction

Krylov subspace methods in finite-dimensional spaces is a mature and deeply studied area [87, 55, 90, 33, 25, 41, 85]. Comparatively, this subject has received less attention in the infinite-dimensional setting, particularly in more recent years. Currently, there are several classical papers on the analysis of the convergence properties of these methods in the infinite-dimensional setting on real and complex Hilbert spaces. The background provided here is intended to be a brief overview of the most popular standard Krylov methods along with their most pertinent features, rather than an exhaustive discussion of all the different methods and algorithms. The context of the methods and analysis discussed in this Chapter refers to the *solvable* inverse linear problem

$$Af = g\,, \quad g \in \mathrm{ran}A\,, \tag{2.1}$$

where $A \in \mathscr{B}(\mathcal{H})$ for $\mathcal{H}$ an infinite-dimensional Hilbert space, and $f \in \mathcal{H}$ a solution to (2.1). Discussion of the more recent rational Krylov methods occurs in Chapter 5, in context with the theory presented therein.

The foundations of many of these methods, as found in the classical papers [54, 5, 45, 72, 85, 86] for example, are formulated in the finite-dimensional setting. In discussing Krylov subspace methods in infinite-dimensions, some formulations in setting up the methods in finite-dimensions may be adapted to the appropriate infinite-dimensional complex Hilbert space setting with minor modifications. Similarly, some of the convergence theory may also be suitably generalised to the infinite-dimensional setting. This background begins with some practical aspects, and descriptions of relevant algorithms. Then some of the most popular and well-known Krylov techniques, along with their salient features, are discussed. These aspects are discussed in regard to the linear inverse problem (2.1), unless explicitly noted.

At this level, an informal notion of a Krylov subspace, or cyclic vector space, is introduced to facilitate the discussion in this review; along with the error and residual term. These notions will be repeated and made more formal in the following chapters.

Definition 2.1.1. The N-th order Krylov subspace with respect to the operator A and some $g \in \mathcal{H}$ is

$$\mathcal{K}_N\left(A,\, g\right) := \operatorname{span}\left\{A^n g \,\middle|\, n \in \{0, 1, \ldots, N-1\}\right\} , \tag{2.2}$$

with a (possibly) infinite-dimensional counterpart

$$\mathcal{K}\left(A,\, g\right) := \lim_{N\to\infty} \mathcal{K}_N\left(A,\, g\right) . \tag{2.3}$$

Informally, these subspaces are referred to in this review as ‘the Krylov (sub)space’ where no confusion arises.

For the numerical approximate, or ‘iterate’, at the N-th step of a method, the preferred notation for this vector is $f^{[N]} \in \mathcal{H}$; with associated residual

$$\mathfrak{R}_N := g - Af^{[N]} . \tag{2.4}$$

Under the condition that (2.1) has a unique solution f, the error is

$$\mathscr{E}_N := f - f^{[N]} . \tag{2.5}$$

2.2 Some orthornomalisation algorithms

Over many years there have several orthonormalisation algorithms that have been developed. Some of these algorithms have been developed with specific operator classes in mind (e.g., self-adjoint operators), all with the goal to find an orthonormal basis for the Krylov subspace in a numerically stable way. Informally speaking, a numerically stable procedure is one whose output is not significantly affected by small errors in inputs. Two of the most popular methods are presented within this Section, namely the Arnoldi and Lanczos algorithms, and how they approximate the infinite-dimensional operator A as a finite-dimensional one. At this point, it should be stressed that these algorithms themselves only construct the Krylov *'search space'* (i.e., the space in which one searches for solution(s)) for the Krylov projection method. A brief survey of these algorithms, and some methods that use them, may be found in [87, 55, 33].

Arnoldi algorithm

This algorithm, as first proposed by Arnoldi [5], is a modified Gram-Schmidt process, and it is the algorithm underpinning the procedure used in the GMRES method [86]. The algorithm is formulated so that it is numerically stable as the iterations proceed. Generally speaking, this algorithm factorises the operator to a matrix, known as the upper Hessenberg matrix, with non-zero entries in the first 'sub-diagonal'. The Arnoldi algorithm (Algorithm 1) works by taking an initial vector $u \in \mathcal{H}$ and finds and orthonormal basis for $\mathcal{K}_N\left(A,\, u\right)$.

The resulting upper Hessenberg matrix is

$$H_{N+1,N} = \begin{pmatrix} h_{1,1} & h_{1,2} & h_{1,3} & \cdots & h_{1,N} \\ h_{2,1} & h_{2,2} & h_{2,3} & \cdots & h_{2,N} \\ 0 & h_{3,2} & h_{3,3} & \cdots & h_{3,N} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & h_{N-1,N} & h_{N,N} \\ 0 & 0 & 0 & 0 & h_{N+1,N} \end{pmatrix}, \tag{2.6}$$

Algorithm 1: Arnoldi Algorithm [55]

```
Data: Operator $A$, vector $u$.
Result: At each step $N$: orthonormal basis vectors $v_1, \ldots, v_N$ for
        $\mathcal{K}_N(A, u)$, Hessenberg matrix $H_{N+1,N}$.
Initialisation: $v_1 = u / \|u\|_{\mathcal{H}}$;
while $\mathcal{K}_N(A, u) \subsetneq \mathcal{K}_{N+1}(A, u)$ do
    $\tilde{v} = Av_N - \sum_{i=1}^{N} h_{i,N} v_i$ where $h_{i,N} = \langle v_i, Av_N \rangle$;
    $h_{N+1,N} = \|\tilde{v}\|_{\mathcal{H}}$;
    if $h_{N+1,N} = 0$ then
        break;
    else
        $v_{N+1} = \tilde{v} / h_{N+1,N}$;
    end
end
```

where $H_{N+1,N} \in \mathbb{C}^{(N+1)\times N}$. One may see that the algorithm terminates if $h_{N+1,N} = 0$, so that $\mathcal{K}_{N+1}(A, u) = \mathcal{K}_N(A, u)$ indicating that the N-th order Krylov subspace is invariant under the action of A. Algorithm 1 gives the following factorisation of the operator A.

$$AV_N = V_{N+1} H_{N+1,N} \,, \tag{2.7}$$

where the partial isometry $V_N : \mathbb{C}^N \to \mathcal{H}$ is constructed using the orthonormal basis vectors found from the algorithm for $\mathcal{K}_N(A, u)$ as its columns, $V_N = \begin{pmatrix} v_1 & v_2 & \cdots & v_N \end{pmatrix}$. This explicitly reveals that the upper Hessenberg matrix is a factorisation that may be seen as a projection of A onto finite-dimensional Krylov subspaces.

Lanczos algorithm

The Lanczos algorithm [54] is only applicable to self-adjoint operators, and is mathematically equivalent to the Arnoldi algorithm in this case. Sometimes the Lanczos algorithm is referred to as the *symmetric* Lanczos algorithm. Here, the orthonormalisation procedure simplifies considerably to a three term recurrence, which is advantageous for numerical calculations. The operator is factorised to a tridiagonal matrix that is perfectly suited to sparse numerical operations. This algorithm, and its variants, underlie the procedures used in conjugate-gradient style methods. The Lanczos algorithm (Algorithm 2) takes an initial vector $u \in \mathcal{H}$ and finds and orthonormal basis for $\mathcal{K}_N\left(A,\, u\right)$.

Algorithm 2: Lanczos Algorithm [55]

```
**Data:** Operator $A$, vector $u$.
**Result:** At each step $N$: orthonormal basis vectors $v_1, \ldots, v_N$ for
        $\mathcal{K}_N(A,\, u)$, tridiagonal matrix $T_{N+1,N}$.
Initialisation: $v_0 = 0$, $\delta_1 = 0$, $v_1 = u/\left\|u\right\|_{\mathcal{H}}$;
**while** $\mathcal{K}_N(A,\, u) \subsetneq \mathcal{K}_{N+1}(A,\, u)$ **do**
    $\tilde{v} = Av_N - \delta_N v_{N-1}$;
    $\hat{v}_{N+1} = \tilde{v} - \gamma_N v_N$ where $\gamma_N = \langle v_N,\, \tilde{v}\rangle$;
    $\delta_{N+1} = \left\|\hat{v}_{N+1}\right\|_{\mathcal{H}}$;
    **if** $\delta_{N+1} = 0$ **then**
        **break**;
    **else**
        $v_{N+1} = \hat{v}_{N+1}/\delta_{N+1}$;
    **end**
**end**
```

At the $(N+1)$-th step, the orthogonality of the vector $\tilde{v}$ to v_N and v_{N-1} is enough to guarantee orthogonality to all the other vectors in $\mathcal{K}_N\left(A,\, u\right)$, as A is self-adjoint [55]. The resulting 'tridiagonal' matrix is

$$T_{N+1,N} = \begin{pmatrix} \gamma_1 & \delta_2 & 0 & \cdots & \cdots & 0 \\ \delta_2 & \gamma_2 & \delta_3 & \ddots & \cdots & 0 \\ 0 & \delta_3 & \gamma_3 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \delta_{N-1} & \gamma_{N-1} & \delta_N \\ 0 & 0 & 0 & 0 & \delta_N & \gamma_N \\ 0 & 0 & 0 & 0 & 0 & \delta_{N+1} \end{pmatrix}, \tag{2.8}$$

where $T_{N+1,N} \in \mathbb{C}^{(N+1)\times N}$. Again, one has the following factorisation of A

$$AV_N = V_{N+1}T_{N+1,N}\,, \tag{2.9}$$

where the partial isometry $V_N : \mathbb{C}^N \to \mathcal{H}$ is constructed using the orthonormal basis vectors found from the algorithm for $\mathcal{K}_N(A, u)$ as its columns, $V_N = \begin{pmatrix} v_1 & v_2 & \cdots & v_N \end{pmatrix}$. Again, this reveals that the tridiagonal matrix is a factorisation that may be seen as a projection of A onto finite-dimensional Krylov subspaces.

2.3 Short review of popular Krylov subspace methods in infinite-dimensions

The issue of convergence of Krylov subspace methods has been a constant topic of research, especially as there are a plethora of different methods available (see [87, 46, 85] for short overviews). Informally speaking, the convergence of an iterative technique is the vanishing of the error term $\mathscr{E}_N$ in some sense (i.e., strongly, weakly, etc), or at the very least the vanishing of the residual $\mathfrak{R}_N$. This topic has been markedly *less* explored in the infinite-dimensional setting, with a large portion of the literature focusing on problems posed in finite-dimensional settings. This Section is devoted to a presentation of the main analysis and results in infinite-dimensions, of course with references to finite-dimensional theory where appropriate.

Historically, most analysis has been devoted to the conjugate-gradient method and its mathematical equivalents, e.g., LSQR [72]. As such, the focus herein remains on the conjugate-gradient method, but also its popular alternative, the generalised minimal residual method (GMRES).

While there is a general convergence theory for conjugate-gradient like methods in infinite-dimensions [64], for the GMRES method there does not appear to be a comparatively general convergence analysis to date. There have been some attempts made for subclasses of operators under particular assumptions that shall be mentioned throughout.

In particular, an attractive playground for convergence analysis is presented by the *fixed point* problem (called as such for when the vector $g = 0$) [66]

$$f = Kf + g\,, \tag{2.10}$$

for $K \in \mathscr{B}(\mathcal{H})$ and $g \in \mathrm{ran}(\mathbb{1} - K)$. Also attractive for convergence analysis is the variation on (2.10)

$$(\zeta\mathbb{1} - K)f = g\,, \quad g \in \mathrm{ran}(\zeta\mathbb{1} - K)\,, \tag{2.11}$$

where $\zeta \in \rho(K)$. Essentially this is just (2.1) where $A = \zeta\mathbb{1} - K$. These problems are attractive for convergence analysis because, in the scenario K is a compact operator on separable Hilbert space, it has a discrete spectrum. The operator A then has a cluster of eigenvalues near the point $\zeta \in \mathbb{C}$, say within a disc centred at $z = \zeta$ with given radius r, and a finite number of points outside the disc. This makes the convergence analysis simpler when using the general functional calculus, as the spectrum is split into these two distinct parts, of which the eigenvalues in the disc are paramount for the asymptotic convergence properties (e.g., see [12]).

2.3.1 Conjugate-gradient methods

The conjugate-gradient method is now a well-studied and understood Krylov subspace method that is historically one of the most significant. It was first introduced by Hestenes and Stiefel [45] and applies to systems (2.1) for *positive* operators, i.e., where $\langle\psi,\, A\psi\rangle \geq 0$ for all $\psi \in \mathcal{H}$.

Conjugate-gradient style methods solve minimisation problems using the N-th order Krylov subspace to approximate the solution. An initial guess to the solution is chosen, $f^{[0]}$, and the Krylov subspace associated with the initial residual is built, namely $\mathcal{K}_N\,(A,\, \mathfrak{R}_0)$. Owing to possible non-injectivity of A, the (possibly) affine *solution* manifold to the linear inverse problem $\mathcal{S}(A, g)$ for a given operator A and datum g is defined as follows

$$\mathcal{S}(A, g) := \{f \in \mathcal{H} \,|\, Af = g\}\ . \tag{2.12}$$

This manifold is non-empty (as $g \in \mathrm{ran}A$), *closed* and *convex* (as $\ker A$ is closed and linear). There exists a projection operator $\mathcal{P}_{\mathcal{S}} : \mathcal{H} \to \mathcal{H}$ that maps elements $f \in \mathcal{H}$ to the *unique* element $\tilde{f} \in \mathcal{S}(A, g)$ such that $\tilde{f} = \mathrm{arginf}_{v \in \mathcal{S}(A,g)} \|v - f\|_{\mathcal{H}}$ [10, Chapter 5].

The N-th iterate is then chosen by solving the minimisation in the affine space $\{f^{[0]}\} + \mathcal{K}_N\,(A,\,\mathfrak{R}_0)$

$$f^{[N]} = \underset{h \in \{f^{[0]}\} + \mathcal{K}_N(A,\,\mathfrak{R}_0)}{\operatorname{argmin}} \left\| A^{\frac{\xi}{2}}(h - \mathcal{P}_{\mathcal{S}} f^{[0]}) \right\|_{\mathcal{H}} \tag{2.13}$$

where $\mathcal{P}_{\mathcal{S}} f^{[0]}$ is a solution to the solvable problem (2.1) that is *closest* in $\mathcal{H}$-norm to the initial guess $f^{[0]}$. The cases of practical interest are $\xi = 1$ or 2. $\xi = 1$ corresponds to error minimisation in the 'energy' (semi-)norm, i.e.

$$\|\cdot\|_A^2 := \langle \cdot,\, A\cdot \rangle \ , \tag{2.14}$$

over the affine space $\{f^{[0]}\} + \mathcal{K}_N\,(A,\,\mathfrak{R}_0)$ [55, 87]. $\xi = 2$ corresponds to residual minimisation in the $\mathcal{H}$-norm over $\{f^{[0]}\} + \mathcal{K}_N\,(A,\,\mathfrak{R}_0)$. In fact, when $\xi = 2$, this becomes mathematically equivalent to the MINRES technique of [72] when applied to positive systems. For *the* conjugate-gradient method, $\xi = 1$.

The solution of the minimisation problem (2.13) for the conjugate-gradient method is equivalent to considering the following projection problem at step N [55]

$$Q_N(Af^{[N]} - g) = 0\,, \tag{2.15}$$

where Q_N is the orthogonal projection operator onto the test space $\mathcal{K}_N\,(A,\,\mathfrak{R}_0)$, and $f^{[N]} \in \{f^{[0]}\} + \mathcal{K}_N\,(A,\,\mathfrak{R}_0)$ is a solution to (2.15).

Clearly, the conjugate-gradient algorithm is also a projection method at each step N, not just an iterative procedure. In the case $\xi = 2$, the test space in equation (2.15) for the projection Q_N is $A\mathcal{K}_N\,(A,\,g)$ instead, while the solution space remains $\mathcal{K}_N\,(A,\,g)$. More precise definitions and aspects of projection methods will be discussed in Chapter 3.

Some of the beginnings of the study in infinite-dimensions of conjugate-gradient methods in *real* Hilbert spaces are presented by Karush [50]. In [50] the author uses the Lanczos algorithm to consider the approximation of the eigenvalue and eigenvector, λ and φ respectively, in equations of the type

$$A\varphi = \lambda\varphi\,, \tag{2.16}$$

and the solution $f \in \mathcal{H}$ to the following linear inverse problem

$$(A - \zeta\mathbb{1})f = g\,, \tag{2.17}$$

where ζ is a given number and $g \in \mathcal{H}$ is a given vector. The assumptions in [50] state that A is a compact, self-adjoint operator, and that as $\overline{\mathcal{K}\,(A,\,g)}$ is A invariant it may be considered as a mapping $A : \overline{\mathcal{K}\,(A,\,g)} \to \overline{\mathcal{K}\,(A,\,g)}$.

The study [50] constructed the sequence of approximate eigenvalues $(\lambda_{jN})_{j\leq N}$ in decreasing order, with corresponding eigenvectors $(\varphi_{jN})_{j\leq N}$ as the eigenvalue-eigenvectors of the operator $Q_N A$, where Q_N is the orthogonal projection onto $\mathcal{K}_N\,(A,\,g)$. λ_{jN}, ψ_{jN} are the N-th step approximate to the j-th eigenvalue and eigenvector respectively. In addition, the following assumption on the eigenvectors is made

$$\|\varphi_{jN}\|_{\mathcal{H}} = 1\,, \quad \langle\varphi_{jN},\, g\rangle > 0\,. \tag{2.18}$$

It should be stressed at this point that the work in [50] considers *indefinite* self-adjoint operators on *not necessarily separable* Hilbert spaces. To derive the first convergence result, Karush [50] uses an assumption that the spectrum $\sigma(A)$ may be written as the union of disjoint sets $\sigma_1(A)$ and $\sigma_2(A)$, where $\sigma_1(A)$ contains a *finite* number of *isolated* eigenvalues,

$$\lambda_1 > \lambda_2 > \cdots > \lambda_m$$

for some $m \geq 1$; and $\sigma_2(A)$ contains spectral values less than λ_m. This assumption is true for compact operators in *separable* Hilbert space (e.g. see [51, 77]). Under these conditions [50, Theorem 1] states that for fixed $j \leq m$; $(\lambda_{jN})_{j\leq N}$ is monotonically increasing and $\lambda_{jN} \to \lambda_j$ as $N \to \infty$. Similarly, $\|\varphi_{jN} - \varphi_j\|_{\mathcal{H}} \to 0$ as $N \to \infty$.

The analysis was expanded to include also estimates on the *rate* of the convergence of the eigenvalues of A. The convergence of eigenvalues and eigenvectors for a *fixed* j is faster than any geometric sequence with positive ratio for sufficiently large N [50]. As this result is for a fixed j, it remains unclear if the convergence rate is *uniform* for eigenvalue-eigenvector pairs.

Similar convergence results were derived for the numerical approximants to (2.17), using the method in [54], to the true solution $f \in \mathcal{H}$ under the condition $\zeta \notin \sigma(A)$. In fact, Karush [50] again shows that the rate of convergence of the error is faster than any geometric sequence. This fact is re-proven in a later study by Daniel [21].

The study [50] thus reveals some of the beginnings of the study of infinite-dimensional theory for Krylov subspace methods, in particular when applied to compact, self-adjoint operators.

Future studies have moved onto the conjugate-gradient method in real Hilbert space. Some earlier works in this direction include [61, 21, 49], while later studies include [97, 64, 65, 9, 56, 44]. The works by [61, 21] are some of the earliest studies that reveal convergence properties of iterative methods in infinite-dimensional Hilbert space. Furthermore, both these studies present a suitably generalised analysis of these methods to non-linear operator equations in the infinite-dimensional setting. Of the two studies by Daniel [21] and Nashed [61], [21] contains a dedicated, in-depth analysis of the conjugate-gradient method.

In the analysis of linear equations of type (2.1), Daniel [21] constructs the convergence theory using the class of positive definite operators $A > \mathbb{O}$ with everywhere defined, bounded inverse on a *real* separable Hilbert space $\mathcal{H}$. Under these conditions, Daniel [21] shows that the sequence of iterates $f^{[N]}$ in (2.13) for conjugate-gradients approaches the true solution f to (2.1) with geometrically fast rate. In fact, the exact convergence formula derived is as follows.

Theorem 2.3.1 (Theorem 1.2.2 [21]). *Consider the linear inverse problem* (2.1) *on a real Hilbert space $\mathcal{H}$ for $A : \mathcal{H} \to \mathcal{H}$ a bounded positive definite operator with everywhere defined bounded inverse. Let $\kappa := (\sup_{\lambda\in\sigma(A)} \lambda)/(\inf_{\lambda\in\sigma(A)} \lambda)$, and $f^{[0]}$ be an initial guess to the solution f. Then* *the error energy functional, as defined in* (2.14)*, decays as follows.*

$$\left\| f^{[N]} - f \right\|_A \leq 2 \left\| f^{[0]} - f \right\|_A \left(\frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \right)^N . \tag{2.19}$$

Moreover, as $\|A^{-1}\|_{\text{op}} < \infty$*, the error also converges at least with the same geometric rate in* (2.19)*.*

This is now a classical and well understood result for conjugate-gradient methods, and has been reproduced in several monographs, e.g., [87]. It has even recently been reformulated, under particular assumptions, in terms of an *unbounded*, strictly positive definite, second order differential operator [35].

Daniel [21] also noted that under particular circumstances, a more rapid rate of convergence may occur. The following theorem from [21] showed that a rate faster than any geometric sequence is possible, much the same as what was proven in [50]. This was later independently investigated and expanded by Winther [97].

Theorem 2.3.2 (Corollary 1 [21])**.** *If there exists some* $\zeta > 0$ *such that* $A - \zeta\mathbb{1}$ *is a compact operator, for* A *as described in Theorem 2.3.1, then the error for the conjugate-gradient method converges faster than any geometric sequence with positive ratio.*

The exact same result was later derived in [97, Remark 2.1]. Under the extra assumption that $A - \mathbb{1}$ is a *p-nuclear* operator (see [77]) with $p \in [1, \infty)$, Winther [97] proves the residual decay rate $\|\mathfrak{R}_N\|_{\mathcal{H}} \leq c_N^N \|\mathfrak{R}_0\|_{\mathcal{H}}$, for $c_N \sim (1/N)^{1/p}$.

To describe better different convergence rates, the following definition is in order.

Definition 2.3.3 ([44])**.** Let $(e_k)_{k \in \mathbb{N}}$ be a sequence of non-negative real numbers converging to zero. Then

(i) The convergence rate is *Q-linear* if there exists some $q \in (0, 1)$ such that $e_{k+1} \leq q e_k$ for $k \geq k_0 \in \mathbb{N}$.

(ii) The convergence rate is *Q-superlinear* if there exists a non-negative sequence $(\varepsilon_k)_{k\in\mathbb{N}}$ converging to zero such that $e_{k+1} \leq \varepsilon_k e_k$ for $k \geq k_0 \in \mathbb{N}$.

(iii) The convergence rate is *R-linear* if $\limsup_{k\to\infty} e_k^{1/k} = r$ for some $r \in (0,1)$.

(iv) The convergence rate is *R-superlinear* if $\lim_{k\to\infty} e_k^{1/k} = 0$.

Informally, one may not distinguish between the 'Q-' and 'R-' prefixes, and just refer to *linear* or *superlinear* convergence. Theorem 2.3.1 therefore contains a result of linear convergence, while Theorem 2.3.2 contains a result for superlinear convergence. In fact, under the the same conditions of Theorem 2.3.2 as described in [97], the convergence is R-superlinear.

The monograph [66] contains a very detailed general analysis of different convergence behaviour in infinite-dimensions for a variety of iterative techniques, with particular emphasis on the fixed point problem (2.10). For a modern survey on linear and superlinear convergence results for the conjugate-gradient and MINRES methods in real Hilbert spaces (also see [44]).

Daniel [21] also studied the conjugate-gradient method applied to non-linear equations. In [21] within the setting of a continuous *non-linear* operator $J : \mathcal{H} \to \mathcal{H}$, $f \mapsto J(f)$ on a real Hilbert space, with a bounded Frechet derivative J'_f of range $\mathcal{H}$, the conjugate-gradient method was suitably modified to solve the equation $J(f) = 0$ under the additional assumptions that J'_f is self-adjoint and coercive (see [21, Section 2.0] for details). Under these conditions, [21, Theorem 2.0.1] shows that the sequence of iterates $f^{[N]}$ generated by the method converges strongly to the *unique* solution f to $J(f) = 0$. Some of these assumptions are then relaxed and assumed to hold only in a convex domain in $\mathcal{H}$ (see [21, Sect. 2.1] for more details). According to Gilles and Townsend [35], these studies by Daniel [21] form some of the first attempts to develop Krylov methods for unbounded differential operators.

A seminal work on conjugate-gradient methods by Kammerer and Nashed [49] was applied to operator equations of the following type

$$Tf = g\,, \quad g \in \mathcal{H}_2 \tag{2.20}$$

for $T : \mathcal{H}_1 \to \mathcal{H}_2$ a bounded linear operator between two *real* Hilbert spaces. In this scenario, Kammerer and Nashed [49] considered finding a solution $f \in \mathcal{H}_1$ to (2.20) as a minimiser of the residual

$$f = \underset{\tilde{f} \in \mathcal{H}_1}{\operatorname{argmin}} \left\| T\tilde{f} - g \right\|_{\mathcal{H}_2} , \tag{2.21}$$

should such a solution exist, known as a 'best approximate' solution. In order solve this system using a conjugate-gradient method, [49] directly worked with $T^*Tf = T^*g$, instead of (2.20). The study includes a preliminary result under the condition of the closed range of T.

Theorem 2.3.4 (Theorem 4.1 [49])**.** *Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces over the real field and let $T : \mathcal{H}_1 \to \mathcal{H}_2$ be a bounded linear operator. If* $\mathrm{ran}T = \overline{\mathrm{ran}T}$*, then the conjugate-gradient method applied to $T^*Tf = T^*g$ (see [49]) converges monotonically to a single best approximate solution f of $Tf = g$. Moreover, if m and M are the greatest lower and least upper spectral bounds, respectively, of the domain restricted operator $T^*T|_{\mathrm{ran}T^*}$, then*

$$\left\| f^{[N]} - f \right\|_{\mathcal{H}_1}^2 \leq \frac{C_0}{m} \left(\frac{M-m}{M+m} \right)^{2N} , \tag{2.22}$$

where C_0 is a constant depending on the initial guess $f^{[0]}$.

This theorem is analogous to Theorem 2.3.1 from [21], however more general as it is not assumed that T^*T has an everywhere defined bounded linear inverse.

Removing the restriction of closed range of T in Theorem 2.3.4, it was again shown that the sequence of errors $(\mathscr{E}_N)_{N\in\mathbb{N}_0}$ monotonically approaches 0 in the strong topology, for f a particular best approximate solution to $Tf = g$ (see [49, Theorem 5.1] for details).

Over a decade later, this work by [49] inspired both Louis [56] and Brakhage [9] to consider the solution to the system (2.20) over *complex* separable Hilbert spaces for T a compact linear operator. In particular [56] built on the work in [9] and found that the convergence rate of the residual and error terms were related to the singular values of T^*T.

Theorem 2.3.5 (Lemma 3.1 and Theorem 3.3 [56])**.** *Consider equation* (2.20) *such that it has a solvable counterpart $T^*Tf = T^*g$ for T compact. Under the conditions that $(T^*T)^\nu u = f$ is solvable for $\nu < 0$, and the initial guess in the conjugate-gradient method $f^{[0]}$ is* 0*, the residual and error terms decay as:*

$$\begin{aligned} \left\|T(f - f^{[N]})\right\|_{\mathcal{H}_2} &\leq \sigma_{N+1}^{-2\nu+1} \left\|R_N(T^*T)^\nu f\right\|_{\mathcal{H}_1} \\ \left\|f - f^{[N]}\right\|_{\mathcal{H}_1} &\leq \sigma_{N+1}^{-2\nu} \left\|R_N(T^*T)^\nu f\right\|_{\mathcal{H}_1}^{-2\nu/(1-2\nu)} \left\|(T^*T)^{2\nu}\right\|_{\mathcal{H}_1}^{1/(1-2\nu)} , \end{aligned} \tag{2.23}$$

where the sequence $(\sigma_n)_{n\in\mathbb{N}}$ are the singular values of T in decreasing order. Here $R_N : \mathcal{H}_1 \to \mathcal{H}_1$ is a bounded linear operator with the property that for any $u \in \mathcal{H}_1$, $\|R_N u\|_{\mathcal{H}_1} \to 0$ as $N \to \infty$, defined by the projection

$$R_N u = \sum_{n=N+1}^{\infty} \langle \varphi_n, u\rangle \varphi_n , \tag{2.24}$$

where $(\varphi_n)_{n\in\mathbb{N}}$ is the orthonormal system of canonical basis vectors for the operator T^*T.

It was observed in [56] that the decay rate presented in Theorem 2.3.5 may not be sharp with respect to the singular value decay rates, due to the extra term involving the operator R_N that also decays. It is worth emphasising that the works by [56, 9, 49] rely on $g \in \operatorname{ran}T \oplus \operatorname{ran}T^\perp$ for their convergence estimates. Aspects of the regularising properties of the conjugate-gradient method are pointed out in the monographs [41, 25] and in the papers [63, 24] among others.

Although the aforementioned studies have significantly built and improved on previous works, perhaps the most *profound* study on the conjugate-gradient method was published in two parts by Nemirovskiy and Polyak [64] and [65], and reproduced in the monographs by Hanke [41] and Engl, Hanke, and Neubauer [25]. These two works [64, 65] definitively showed that for a bounded, self-adjoint, positive operator on *complex* Hilbert space, the numerical approximants from the conjugate-gradient method converge to a single solution to (2.1). Moreover, in [65] the authors showed that it is impossible to improve the convergence rate estimates among the whole class of bounded, self-adjoint, positive operators, where $0 \in \sigma(A)$ is *not* an isolated point. This shows the *optimality* or *sharpness* of the convergence results in [64]. Under the specific assumption of A as a compact operator, the convergence rates stated in [64] are sharper than the results derived in [56]. To date, [64, 65] appear to be the most *general* studies of the convergence properties of the conjugate-gradient method. Although the analysis in [64] contains general constructions and convergence rate estimates for *several* iterative techniques, the focus here remains on the convergence estimates for the conjugate-gradient method.

Nemirovskiy and Polyak [64] begins by considering (2.1) for a self-adjoint, positive operator, possibly with $0 \in \sigma(A)$. Here, $\mathcal{P}_{\mathcal{S}}$ is the projection operator as defined for the solution manifold (2.12).

Theorem 2.3.6 (Theorem 7 [64])**.** *Consider the problem* (2.1) *with* $A = A^*$ *and* $\langle x, Ax \rangle \geq 0$ *for all* $x \in \mathcal{H}$. *Consider the sequence of iterates* $(f^{[N]})_{N \in \mathbb{N}_0}$ *in* $\mathcal{H}$ *defined by* (2.13) *for* $\xi = 1$. *Then one has that*

$$\left\| f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]} \right\|_{\mathcal{H}} \xrightarrow{N \to \infty} 0 \tag{2.25}$$

and moreover, for every $\nu < 0$

$$\left\| f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]} \right\|_{\mathcal{H}} \leq \left(\frac{C_{f^{[0]},\nu}}{2N+1} \right)^{-2\nu}, \tag{2.26}$$

for a constant $C_{f^{[0]},\nu} > 0$ *depending on the initial guess* $f^{[0]}$ *and* $\nu < 0$, *provided that* $A^{-\nu} u = f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]}$ *has a solution* $u \in \mathcal{H}$.

In fact, the original work [64] contains a more general statement, however the above theorem suffices for error convergence rates. Also of interest is the case under the assumption that $\sigma(A)$ is the closure of a decreasing sequence $(\sigma_n)_{n \in \mathbb{N}} \subset \mathbb{R}^+$, such as the case when A is a compact operator as considered in [56]. In this event, the following corollary holds.

Corollary 2.3.7 (Equation 3.13′ [64])**.** *Under the conditions stated in Theorem 2.3.6 in addition with the spectrum of* A *being the closure of a sequence of decreasing positive real numbers* $(\sigma_n)_{n \in \mathbb{N}}$, *the following estimate holds for* *the error in the approximation.*

$$\left\|f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]}\right\|_{\mathcal{H}} \leq C_{f^{[0]},\nu} \min_{0\leq k\leq N} \left\{\sigma_{k+1}^{-2\nu}(N-k+1)^{4\nu}\right\} . \tag{2.27}$$

Corollary 2.3.7 provides a sharper estimate than immediately available from Theorem 2.3.5 provided by [56] for compact operators.

2.3.2 Generalised minimal residual (GMRES) methods

Compared to conjugate-gradients, GMRES was first formulated decades later by Saad and Schultz [86]. It has become a widely used and well-studied Krylov subspace method. Much of the analysis of this method has occurred in the finite-dimensional setting (see [85, 86] for some general overviews). The analysis of this method in the infinite-dimensional setting, particularly on the Krylov solvability of the linear inverse problem, remains elusive, although some studies present results under specific assumptions on the operator [32, 66, 12, 57], and more recently [68, 67]. In particular, the monograph [66] contains a number of approximation results for GMRES residual polynomials, with the particular focus being the fixed point problem. The monograph [53] also contains general remarks and results for minimal residual methods in infinite-dimensions. More recently, the analysis of the GMRES method has been extended to a sub-class of unbounded differential operators [69, 70, 71], known therein as ‘differential GMRES’.

As the conjugate-gradient method applies to a subclass of self-adjoint operators, the tools of the spectral integral are immediately available (see [88, Chapters 4 & 5] or Appendix B for details) and underpin much of the theory presented in Section 2.3.1. Now, the theory of the functional calculus is only available in the algebra $\mathscr{B}(\mathcal{H})$ as a Cauchy integral (see [51] and [76, Chapter XI] for details). In the case where A is self-adjoint, the GMRES method applied to (2.1) is mathematically equivalent to the MINRES method of [72].

The iterates $(f^{[N]})_{N\in\mathbb{N}}$ of the GMRES method are the solutions to the following minimisation problem at each step N, where $f^{[0]}$ is an initial guess,

$$f^{[N]} = \underset{h \in \{f^{[0]}\} + \mathcal{K}_N(A, \mathfrak{R}_0)}{\operatorname{argmin}} \|Ah - g\|_{\mathcal{H}} \,. \tag{2.28}$$

Again, this may be posed in the same form as (2.15), except that Q_N is now the orthogonal projection operator onto the subspace $A\mathcal{K}_N(A, \mathfrak{R}_0)$.

An interesting comment regarding the convergence behaviour of the GMRES method may immediately be made at the *finite*-dimensional level. In a series of articles [37, 38, 4] it was shown that information on the eigenvalues of a matrix is not enough to determine the convergence behaviour of the residual. In fact, in Arioli, Pták, and Strakoš [4] the authors show that the convergence behaviour of the algorithm is independent on the eigenvalues of the matrix system. For infinite-dimensional GMRES systems, heuristically speaking [66, Section 1.8] makes the comment that the convergence curve can '...look like almost any curve by tailor making a strange operator and initial residual.'

Theorem 2.3.8 (Theorem 2.1 [4])**.** *Consider the finite-dimensional Hilbert space $\mathbb{C}^m$ for some $m \in \mathbb{N}$. Suppose that $(\lambda_N)_{N \leq m} \subset \mathbb{C}$ is a set of non-zero numbers, and that $g \in \mathbb{C}^m$ a non-trivial m-dimensional vector. Let $(l_N)_{N \leq m}$ be a monotonically decreasing sequence such that $l_m = 0$. Then, there exists a matrix $A \in \mathbb{C}^{m \times m}$ with eigenvalues $(\lambda_N)_{N \leq m}$ such that the GMRES method applied to the linear inverse problem* (2.1) *generates a sequence of residuals $(\|\mathfrak{R}_N\|_{\mathbb{C}^m})_{N \leq m}$ such that $\|\mathfrak{R}_N\|_{\mathbb{C}^m} = l_N$.*

General convergence analysis of the GMRES method in finite-dimensions may be found in [86, 96, 11, 73, 12], and in particular the monograph [87, Sect. 6.11.4]. The case when the operator A is a diagonalisable matrix is of particular interest as follows.

Theorem 2.3.9 (Corollary 6.1 [87])**.** *Consider the linear inverse problem* (2.1) *in $\mathbb{C}^m$ for m finite and equipped with the standard dot scalar product. Let $A \in \mathbb{C}^{m \times m}$ be a diagonalisable matrix $A = X\Lambda X^{-1}$, where $\Lambda = \operatorname{diag}\{\lambda_1, \lambda_2, \ldots, \lambda_m\}$ is the matrix of eigenvalues. Assume that all the eigenvalues of A are located in an ellipse $E(c, d, a)$, centred at c with focal* *distance d and semi-major axis a, that excludes the origin. Then the residual norm achieved at the N-th step of GMRES satisfies the inequality*

$$\|\mathfrak{R}_N\|_{\mathbb{C}^m} \leq \|X\|_{\mathrm{op}} \left\|X^{-1}\right\|_{\mathrm{op}} \left(\frac{C_N(a/d)}{C_N(c/d)}\right) \|\mathfrak{R}_0\|_{\mathbb{C}^m} \ , \tag{2.29}$$

where $C_k(z) = (1/2)(w^k + w^{-k})$ and $z = (1/2)(w + w^{-1})$.

The analysis available in the published literature on the GMRES method in the infinite-dimensional setting is also restricted to certain operator classes. Of course, general aspects of convergence and polynomial methods may be found (e.g. [66]), and some attempts to characterise the Krylov solvability of (2.1) and convergence behaviour of the GMRES method [12, 13, 32, 69, 57] have been made. The focus is primarily on the *convergence behaviour* of GMRES rather than operator-theoretic notions of Krylov solvability.

In two papers, [12] and [13], the GMRES method was analysed in infinite-dimensional Hilbert space. Under some assumptions on the structure of the operator A, the solution to (2.1) using GMRES was shown to exhibit superlinear convergence in the residual. The study [12] first considered the situation under which A is in the finite-dimensional setting, and then extended the result to the infinite-dimensional setting. The finite-dimensional results in [12] derive superlinear convergence rates by considering clustering of the eigenvalues of the operator A, and the minimal polynomial of the eigenvalues outside the cluster. This is analogous to the proof in [50] for self-adjoint operators.

For completeness it is appropriate to define the minimal polynomial at the present moment, in particular as it is a tacitly used concept at the finite-dimensional level.

Definition 2.3.10 (Definition 2.8.1 and 2.8.4 [66])**.** A polynomial $p : \mathbb{C} \to \mathbb{C}$

$$p(z) = z^n + \alpha_1 z^{n-1} + \cdots + \alpha_n$$

is called a *minimal* polynomial for the bounded linear operator $A : \mathcal{H} \to \mathcal{H}$ if $p(A) = 0$ and $\tilde{p}(A) \neq 0$ for any non-trivial polynomial $\tilde{p}$ of lower degree than p. An operator is said to be *algebraic* of degree n if it has a minimal polynomial of degree n.

The extension of the finite-dimensional analysis by Campbell et al. [12] leads to the following proposition that shows the R-superlinear convergence of GMRES.

Proposition 2.3.11 (Proposition 6.1 [12]). *Consider the linear inverse problem* (2.1) *such that $A = \mathbb{1} + K$ has $0 \in \rho(A)$ and K is a compact linear operator on $\mathcal{H}$. Consider some $\alpha > 0$ and define the 'cluster' to be the set $\{z \in \mathbb{C};\, |z-1| < \alpha\}$. Consider the finite collection of outlier eigenvalues $(\lambda_j)_{1\le j\le M}$ of A, such that they do* not *belong to the cluster. Define the 'distance' between the outliers and the cluster as*

$$\delta := \max_{|z-1|=\alpha}\ \max_{1\le j\le M} \frac{|\lambda_j - z|}{|\lambda_j|}\,.$$

Then for any $g \in \mathcal{H}$ and any $f^{[0]}$, one has

$$\|\mathfrak{R}_{d+k}\|_{\mathcal{H}} \le C\alpha^k \|\mathfrak{R}_0\|_{\mathcal{H}}\ , \tag{2.30}$$

where d is the degree of the minimal polynomial of the outliers, and C is a constant dependent on d, and δ and is independent of k.

Proposition 2.3.11 shows that $\limsup_{k\to\infty} \|\mathfrak{R}_k\|_{\mathcal{H}}^{1/k} \le \alpha$. As $\alpha > 0$ is arbitrary, this implies that $\lim_{k\to\infty} \|\mathfrak{R}_k\|_{\mathcal{H}}^{1/k} = 0$ so that the convergence is R-superlinear.

This result from [12] shows analogous conditions and behaviour to the previously mentioned results of [50, 21, 97] for Lanczos or conjugate-gradient based methods. These types of results are well known under assumptions on A as in Proposition 2.3.11. Heuristically speaking, the GMRES method initially 'sees' the eigenvalue outliers, and after finitely many iterations 'removes' this area of the spectrum [66]. Then the algorithm proceeds to investigate the tight cluster, where the convergence is faster. This can be seen very clearly in the analysis contained in [12] as well as numerous comments and results presented [66].

In a related follow-up Campbell et al. [13] considered similar operators as in Proposition 2.3.11. In this setting, however, the authors also considered the discretisation of the method. It was then shown that the performance of the GMRES method becomes independent of the resolution of the discretisation if it is sufficiently fine. In fact the main theorem in [13] is a similar statement to Proposition 2.3.11, except that as under stated extra assumptions, the convergence of the GMRES method is R-superlinear and independent of the resolution of discretisation [13, Theorem 1.1].

Proposition 2.3.11 from [12], as well as the results on discretisation independence in [13], are in *stark* contrast to Theorem 2.3.8. The studies [12, 13] reveal that the infinite-dimensional nature of the problem gives an asymptotic result that is clearly not accessible at the finite-dimensional level. The conclusions reached in [4] and presented in Theorem 2.3.8, especially about the independence on the eigenvalue distribution, are therefore strictly limited to the finite-dimensional setting.

A short note on the GMRES method by Moret [57] found that in some instances the convergence is Q-superlinear. Moret [57] considered the linear inverse problem (2.1) where $A = \zeta\mathbb{1}+K$ with $0 \in \rho(A)$, and where $K : \mathcal{H} \to \mathcal{H}$ is a compact linear operator and $\zeta \neq 0$. Under these conditions the residuals generated from the algorithm converge Q-superlinearly [57, Theorem 1]. In addition to previous works on the issue, in [57] it was found that the residual convergence rate is related to the product of the singular values of the operator K.

More recently, [32] considered the inverse linear problem (2.1) under the conditions that A is an *algebraic* operator. It is well known that the assumption of an operator being algebraic at the finite-dimensional level is trivial, but this is not so at the infinite-dimensional level [66]. In any case, this assumption immediately guarantees the Krylov solvability in the case that $g \in \operatorname{ran}A$. The authors investigate the circumstances under which the Arnoldi algorithm may break down as well as cases where A is *not* injective, and $g \notin \operatorname{ran}A$. A result on Krylov solvability is stated in the following theorem.

Theorem 2.3.12 (Theorem 4.1 [32]). *Let $A \in \mathscr{B}(\mathcal{H})$ be an algebraic operator of degree $n_0 > 0$. Then the inverse linear problem* (2.1) *(where $g \in \operatorname{ran}A$) has* *a solution in the affine space* $\{f^{[0]}\} + \mathcal{K}_{n_0}(A,\, \mathfrak{R}_0)$ *if and only if* $g \in \mathrm{ran} A^{n_0}$*. In this case, the Krylov solution is unique, and GMRES computes this unique solution.*

Under the case where $g \notin \mathrm{ran} A$, it was shown in [32] that under some restrictions, the GMRES algorithm is capable of calculating a 'least squares solution' [32, Theorem 3.1]. A 'least squares solution' being an approximate solution $\widetilde{f} \in \mathcal{H}$ to the linear inverse problem that minimises $\left\| A^* g - A^* A \widetilde{f} \right\|_{\mathcal{H}}$.

2.4 Further remarks

Even from this relatively short review, one can see that Krylov based methods in infinite-dimensional Hilbert spaces has a history in numerical analysis dating back to the 1950s when various algorithms were initially proposed. There are attractive results for particular classes of problems, for example fixed point and similar problems as in equations (2.10) and (2.11) respectively. The convergence behaviour of both the conjugate-gradient and its related methods is quite well documented, particularly due to the general analysis contained in [64, 65]. The GMRES method on the other hand exhibits rigorously proven convergence behaviour, but the focus is mainly on operators with structure $A = \zeta \mathbb{1} + K$, K is compact and $\zeta \in \rho(A)$. This allows one to focus on the relationship between convergence and the eigenvalues in a disc of arbitrary radius about ζ. Presently, there does not appear to be a general discussion of under what conditions, or exactly what qualifies when a solution is approximable by this method. More generally speaking, a systematic study of when there exists a solution to (2.1) in the Krylov subspace, is missing at the level of closed (possibly unbounded) operators.

Chapter 3

Truncation and Convergence Issues in Hilbert Space

3.1 Introduction

Scientific computing demands a suitable discretisation of the linear inverse problem (1.1). There are certain features, discussed within this Chapter, that appear at the infinite-dimensional level which may become absent at the finite-dimensional level, and yet are relevant to the problem. Of an immediate interest is how close the solution(s) to the discretised problem are to the exact solution(s) of the linear inverse problem. This Chapter, based on the paper [17], focuses on the behaviours of sequences of the appropriately truncated, finite-dimensional problem from the discretisation of the original infinite-dimensional problem.

Recall that the infinite-dimensional linear inverse problem is the problem, given a linear operator $A \in \mathscr{B}(\mathcal{H})$ acting on a Hilbert space $\mathcal{H}$ and some $g \in \mathcal{H}$, to find the solution(s) $f \in \mathcal{H}$ to the linear equation

$$Af = g\,. \tag{3.1}$$

If $g \in \mathrm{ran}A$, then (3.1) is called *solvable*; if in addition A is injective, then (3.1) is called *well-defined*; and finally if in addition, the solution $f \in \mathcal{H}$ is continuously dependent on the datum g, then the problem is *well-posed.*

For certain classes of infinite-dimensional linear inverse problems the theoretical aspects of truncation and convergence issues are already well-established (see, for example [28, 20, 75, 74, 53]). Here, the discussion turns to *generic* truncations of the linear inverse problem, in the sense of *general* projection methods. Under the framework of '*general projection methods*', some relaxations of the well-established theory are made to encompass a broader set of algorithms and solution techniques. The aim of this Chapter is not to develop a comprehensive classification of theoretical and practical phenomena and difficulties occurring from such methods, but to discuss some of the more general features that are unavoidable at a high level of generality. Several model examples, and counter-examples, are developed that illustrate some of the theoretical concepts and challenge the common intuition.

In particular, the errors and residuals arising from the solution of a successively truncated linear inverse problem are discussed. A view to controlling these quantities and their convergence is developed in a *weaker* sense rather than the typical strong topological norm convergence in the Hilbert space. At an abstract level, sufficient conditions are developed that guarantee the error or residual to be small in this more generalised sense.

3.2 Definitions and comments

To begin with, certain notations are laid out to maintain clarity in the following work, and avoid unnecessary confusion. Afterwards, the *exact* framework of the truncations that will be considered is described, namely *general projection methods.*

The following terminology concerns the formalism that is used within this thesis when studying suitably truncated, or discretised, linear inverse problems. One considers the orthonormal systems $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$. These collections need *not* form a basis for the entire space $\mathcal{H}$, however in doing so it is desirable for the 'goodness' of the approximation of the linear inverse problem. Practically speaking, the two sets $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are a-priori known sets of orthonormal vectors that are to be used in the numerical truncation, or discretisation. This is unlike, for example, the singular value decomposition of a compact operator on $\mathcal{H}$, as the basis vectors may not be explicitly known.

A choice regarding $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ is dependent on the numerical scheme chosen. For example, in finite element methods these may be taken as a space of orthogonal, piecewise linear elements on a mesh [28, Chapter 1]. In Krylov subspace methods they are taken from the spanning vectors of the Krylov subspace [55, Chapter 2].

Definition 3.2.1. Let $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be two orthonormal systems of vectors. Given some $N \in \mathbb{N}$, then the orthonormal projections in $\mathcal{H}$ onto $\operatorname{span}\{u_1, \ldots, u_N\}$ and $\operatorname{span}\{v_1, \ldots, v_N\}$ are defined as

$$P_N := \sum_{n=1}^{N} |u_n\rangle \langle u_n| \, , \quad Q_N := \sum_{n=1}^{N} |v_n\rangle \langle v_n| \tag{3.2}$$

respectively.

The framework of the concept of *general projection methods* will be defined in what follows. Herein the definition is given for the class of *bounded* linear operators on Hilbert space. Of course, with minor modifications, the definition may be suitably modified to encompass general unbounded linear operators on Banach spaces (see [53] for the appropriate generalisation of the Galerkin and Petrov-Galerkin methods).

Definition 3.2.2. Let $A : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator on Hilbert space $\mathcal{H}$ and consider the solvable linear inverse problem (3.1). Consider two sequences $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ of orthonormal vectors in $\mathcal{H}$, with associated projections given by (3.2), and the approximation of the solvable problem (3.1) by the following linear inverse problem

$$Q_N A P_N \widehat{f^{(N)}} = Q_N g \, . \tag{3.3}$$

Then the construction of a sequence of solutions to (3.3), $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$, or approximations $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$ to (3.3) in the sense $Q_N A P_N \widehat{f^{(N)}} = Q_N g + \widehat{\varepsilon^{(N)}}$, for some discrepancy $\widehat{\varepsilon^{(N)}} \in \mathcal{H}$, is known as the *general projection method.* In the case where $u_n = v_n$ for all $n \in \mathbb{N}$ the projection method is known as an *othogonal* projection method, otherwise it is termed an *oblique* projection method.

Remark 3.2.3. In Definition 3.2.2, it is *not* guaranteed that (3.3) is solvable. As such, the general projection method permits one to relax assumptions on the solvability of (3.3) and search for *approximate* solutions to the problem instead. This notion is again formulated in the following Section (see (3.12)) at the truncated level.

In (3.3), $Q_N g = \sum_{n=1}^{N} \langle v_n, g \rangle v_n$ is the datum and $\widehat{f^{(N)}} = \sum_{n=1}^{N} \left\langle u_n, \widehat{f^{(N)}} \right\rangle u_n$ is the unknown. The *compression* $Q_N A P_N$ is only non-trivial as a map from $P_N \mathcal{H}$ to $Q_N \mathcal{H}$. The kernel space of the compression $Q_N A P_N$ contains at least the infinite-dimensional space $(\mathbb{1} - P_N)\mathcal{H}$. Normally one has that $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are orthonormal *bases* of the Hilbert space $\mathcal{H}$. In the case of $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ being bases for the Hilbert space $\mathcal{H}$, this is known as the *approximability condition* [28, Definition 2.14].

Clearly, the general projected problem (3.3) is a truncation, or discretisation, of the linear inverse problem (3.1). For convenience of computation, this is cast in terms of a finite-dimensional matrix system on $\mathbb{C}^N$ in what follows.

The finite-dimensional spaces $P_N \mathcal{H}$ and $Q_N \mathcal{H}$, contained in the ambient Hilbert space $\mathcal{H}$, are identified with $\mathbb{C}^N$, i.e., $P_N \mathcal{H} \cong \mathbb{C}^N \cong Q_N \mathcal{H}$. In this way, the vectors $P_N f \in \mathcal{H}$ and $Q_N g \in \mathcal{H}$ are canonically identified with the following finite-dimensional vectors in $\mathbb{C}^N$

$$f_N = \begin{pmatrix} \langle u_1, f \rangle \\ \vdots \\ \langle u_N, f \rangle \end{pmatrix} \in \mathbb{C}^N, \quad g_N = \begin{pmatrix} \langle v_1, g \rangle \\ \vdots \\ \langle v_N, g \rangle \end{pmatrix} \in \mathbb{C}^N. \tag{3.4}$$

From the above description, one now has that the compression $Q_N A P_N$ on Hilbert space $\mathcal{H}$ is identified with a $\mathbb{C}^N \to \mathbb{C}^N$ linear map represented by the matrix $A_N \in \mathbb{C}^{N\times N}$ whose i, j-th entries are given by

$$A_{N;ij} = \langle v_i, Q_N A P_N u_j \rangle . \tag{3.5}$$

The inverse linear problem

$$A_N f^{(N)} = g_N \tag{3.6}$$

with datum $g_N \in \mathbb{C}^N$, unknown $f^{(N)} \in \mathbb{C}^N$, and matrix A_N described by (3.5) is referred to as the N-*dimensional truncation* of the original problem $Af = g$.

At this point, the meaning of the notation is stressed to avoid possible confusion.

(i) $Q_N A P_N$, $P_N f$ and $Q_N g$ are all objects referred to in the whole Hilbert space $\mathcal{H}$, whereas A_N, $f^{(N)}$, f_N and g_N are the analogues referred to in the space $\mathbb{C}^N$.

(ii) The subscript N in the objects A_N, f_N, and g_N indicate that that the components of these objects are precisely the corresponding components (up to order N) respectively of A, f, and g, with respect to the declared orthonormal systems $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$, via (3.4) and (3.5).

(iii) The superscript (N) in $f^{(N)}$ indicates that the components of the $\mathbb{C}^N$-vector $f^{(N)}$ are *not* necessarily understood to be the first N components of the $\mathcal{H}$-vector f with respect to the system $(u_n)_{n\in\mathbb{N}}$. In particular, for $N_1 < N_2$, the components of $f^{(N_1)}$ are not necessarily equal to the first N_1 components of $f^{(N_2)}$. From counterexamples, it is known that if $f \in \mathcal{H}$ is a solution to $Af = g$, in general the truncations A_N, f_N and g_N do *not* satisfy the identity $A_N f_N = g_N$. Therefore, it is essential that the notation $f^{(N)}$ is used for the C^N-vector representation of the unknown in (3.6).

(iv) Lastly, for the $\mathbb{C}^N$-vector $f^{(N)}$, the notation with the wide-hat symbol, $\widehat{f^{(N)}}$, indicates a vector in $\mathcal{H}$ whose first N components with respect to the orthonormal system $(u_n)_{n\in\mathbb{N}}$ are precisely those of $f^{(N)}$, such that $\widehat{f^{(N)}}$ has no vector support in the space span $\{u_1, \dots, u_N\}^\perp$. Therefore $f^{(N)} = (\widehat{f^{(N)}})_N$ and $f_N = (\widehat{f_N})_N$. In general $f \neq \widehat{f_N}$.

In the following, two important subclasses of general projection methods are described, namely the *Galerkin* and *Petrov-Galerkin* methods [53].

Definition 3.2.4. Let $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be two orthonormal bases in the Hilbert space $\mathcal{H}$, and define the projection operators P_N and Q_N as in (3.2). Consider the approximation of the well-defined linear system (3.1) as follows.

$$Q_N(A\widehat{f^{(N)}} - g) = 0\,, \tag{3.7}$$

where $\widehat{f^{(N)}}$ is the unique solution to the well-defined system (3.7) in the space $P_N\mathcal{H}$. Then the construction of the sequence of approximate solutions $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$ to the inverse problem $Af = g$ using (3.7), along with suitable assumptions on A and the truncation bases to guarantee norm convergence $\widehat{f^{(N)}} \xrightarrow{N\to\infty} f$, is called the *Petrov-Galerkin* projection method. Moreover, if $u_n = v_n$ for all $n \in \mathbb{N}$, then this is known as the *Galerkin* projection method.

In this definition of Petrov-Galerkin methods, there exists a unique solution $\widehat{f^{(N)}} \in P_N\mathcal{H}$ for the projected problem at every step N, and both $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are *complete* orthonormal systems in $\mathcal{H}$. Moreover, there are extra conditions that guarantee the strong convergence of the sequence of numerical approximants to the solution of the linear inverse problem $Af = g$. These assumptions are *not* considered in the general projection method definition. In the standard Petrov-Galerkin nomenclature, the u_n's and v_n's span the *solution space* (or *search space*) and the *trial space*, respectively [28, 74].

The framework of Galerkin or Petrov-Galerkin methods is already very well-studied and a classical area of numerical analysis, especially in finite element methods [28, 20]. For certain classes of boundary value problems on $L^2(\Omega)$ for some domain $\Omega \subset \mathbb{R}^n$, the properties and convergence results of the error for these methods are well-established. In these cases, the operator A is of differential type, hence unbounded on $L^2(\Omega)$, but also typically *elliptic* [28, Chapter 3], [74, Chapter 4], or also possibly *Friedrichs* type [28, Section 5.2], [3, 2, 29], or *parabolic* type [28, Chapter 6], [74, Chapter 5], etc. Hence a typical assumption on such operators is that they satisfy some additional coercivity condition, or possibly some other condition among various classical ones that ensure (3.1) is well-posed (e.g., the Banach-Nečas-Babuška Theorem). Usually, the *discretised* system too satisfies the Lax-Milgram lemma, or more generally the Banach-Nečas-Babuška Theorem, at the discrete level; or has assumptions guaranteeing the solution to the discretised finite-dimensional linear inverse problem [28, 75].

Again, to stress the contrast, this is *not* assumed in the general projection framework studied in this Chapter. It may well be the case that the linear inverse problem at the finite-dimensional discretised level does *not* have a solution at a given step N in the discussions that follow. Moreover, the assumption of the density of the search and test spaces in the ambient Hilbert space are not generally assumed to hold. That is, to say, that the 'approximability' of $\mathcal{H}$ need not hold in general. An example would be the case of a Krylov projection method where the underlying Krylov subspace, i.e., the search space, is *not* dense in $\mathcal{H}$. In summary, the theory in this Chapter moves outside the Petrov-Galerkin framework in the sense that

(i) (3.1) is only considered to be solvable,

(ii) generally, the orthonormal systems $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are not assumed to form bases of $\mathcal{H}$,

(iii) the truncated problem (3.3) is not guaranteed to be well-defined, let alone solvable,

(iv) additional assumptions guaranteeing the strong vanishing of the error and/or the residual (see Definition 3.2.5) are not assumed a-priori.

Definition 3.2.5. Consider the linear inverse problem (3.1), and a sequence of approximants $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$ in the ambient Hilbert space $\mathcal{H}$, to a solution $f \in \mathcal{H}$ of (3.1). The infinite-dimensional *error* of the approximation at the N-th step is defined as

$$\mathscr{E}_N := f - \widehat{f^{(N)}}\,, \tag{3.8}$$

and the infinite-dimensional *residual* of the approximation at the N-th step is defined as

$$\mathfrak{R}_N := g - A\widehat{f^{(N)}}\,. \tag{3.9}$$

When no confusion arises, the additional term of 'infinite-dimensional' is dropped when describing error and residual terms. The term 'infinite-dimensional' is only used to distinguish these quantities from the error and residual terms for the finite-dimensional truncated system at fixed N, which may be indexed by the number of steps performed in an iterative algorithm to solve said truncated problem.

3.3 Finite-dimensional truncations

3.3.1 Singularity of the truncated problem

The question of the singularity of the truncated problem (3.3) makes sense *eventually in* N, i.e., for all N's large enough. Certainly, for a *fixed* value of N, the truncation may alter the problem such as to make it uninformative compared to $Af = g$, with such an aberration disappearing for larger values of N.

But even when the solvability of $A_N f^{(N)} = g_N$ is inquired eventually in N, the answer is generally negative as the following simple example shows.

Example 3.3.1. The matrix A_N may remain singular for all $N \in \mathbb{N}$, even for an *injective* operator. For example, consider the weighted (compact) right-shift operator (see Appendix A) on $\ell^2(\mathbb{N})$, and truncate with respect to the canonical basis $(e_n)_{n\in\mathbb{N}}$. Indeed,

$$\mathcal{R}_N = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ \sigma_1 & 0 & \cdots & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_{N-1} & 0 \end{pmatrix} \tag{3.10}$$

is *always* singular, irrespective of N, with $\ker \mathcal{R}_N = \operatorname{span}\{e_N\}$.

Variations on this example where the matrix A_N is alternating between singular and non-singular as $N \to \infty$ are not difficult to construct. If one were to consider, for example, the following operator on $\ell^2(\mathbb{N})$, $A = \sum_{n\in\mathbb{N}} (|e_{n+1}\rangle \langle e_n| + |e_{2n}\rangle \langle e_{2n}|)$, then this would do the trick.

Remark 3.3.2. It is worth noting that when one considers the weighted right-shift $\mathcal{R}$, now over $\ell^2(\mathbb{Z})$, with weights of all of unit value, a truncation scheme using the subset of canonical basis vectors $(e_n)_{-N\leq n\leq N}$ of $\ell^2(\mathbb{Z})$ produces a phenomenon known as *spectral pollution*. This is the occurrence of erroneous eigenvalues at the truncated level that do not converge to any point in the spectrum in the limit $N \to \infty$. This is equivalent to saying that the sequence of truncated operators does not converge to the operator in the *resolvent sense* (see [51, Chapter IV]). Indeed, the only spectral point in the truncated system is 0 for all $N \in \mathbb{N}$, while the spectrum of $\mathcal{R}$ is the unit circle.

Example 3.3.3. On the other hand, it may well be the case that the truncated matrix A_N is always non-singular. For example, the truncation, with respect to the canonical basis $(e_n)_{n\in\mathbb{N}}$ of the multiplication operator on $\ell^2(\mathbb{N})$ with weights $\sigma_n = 1/n$ (see Appendix A) gives

$$M_N = \begin{pmatrix} 1 & \cdots & \cdots & \cdots & 0 \\ 0 & \frac{1}{2} & \cdots & \cdots & 0 \\ 0 & 0 & \frac{1}{3} & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \frac{1}{N} \end{pmatrix}, \tag{3.11}$$

which is a $\mathbb{C}^N \to \mathbb{C}^N$ bijection for all $N \in \mathbb{N}$.

For general projection methods, choosing 'bad' truncations is always possible as the following lemma shows.

Lemma 3.3.4. *Let $\mathcal{H}$ be a separable Hilbert space with* $\dim \mathcal{H} = \infty$, *and let $A \in \mathscr{B}(\mathcal{H})$. There always exist two orthonormal systems $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ of $\mathcal{H}$ such that the corresponding truncated matrix A_N defined as in* (3.5) *is singular for every $N \in \mathbb{N}$.*

Proof. Pick an arbitrary orthonormal basis (or less restrictively, an orthonormal system) $(u_n)_{n\in\mathbb{N}}$ so that one may then construct the other orthonormal system $(v_n)_{n\in\mathbb{N}}$ inductively.

When $N=1$, it suffices to choose v_1 such that $v_1 \perp Au_1$ and $\|v_1\|_{\mathcal{H}}=1$. Now let $(v_n)_{n\in\{1,\dots,N-1\}}$ be an orthonormal system in $\mathcal{H}$ satisfying the thesis up to the order $N-1$. Then there exists a choice of v_N such that the final row of the matrix A_N has all zero entries, as will now be shown. In fact, $(A_N)_{ij} = (Q_N A P_N)_{ij} = \langle v_i, Au_j\rangle$ for $i\in\{1,\dots,N-1\}$ and $j\in\{1,\dots,N\}$. In order for $\langle v_N, Au_j\rangle = 0$ for $j\in\{1,\dots,N\}$, it suffices to take

$$v_N \perp \mathrm{ran}(AP_N)\,,\quad v_N \perp \mathrm{ran}Q_{N-1}\,,\quad \|v_N\|_{\mathcal{H}}=1\,,$$

where P_N and Q_{N-1} are the orthogonal projections defined by (3.2). Since the spaces $\mathrm{ran}(AP_N)$ and $\mathrm{ran}Q_{N-1}$ are finite-dimensional subspaces of $\mathcal{H}$, there must exist a vector $v_N\in\mathcal{H}$ with the above stated properties. □

Although this lemma is stated for general projection methods, Example 3.3.1 shows a typical case of this occurring under the condition that $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ are orthonormal *bases* of $\mathcal{H}$ (and moreover $u_n=v_n$). Surely then, the question of when bad truncations occur is not so 'trivial' a question to be asked at such a high level of generality. In the standard framework of (Petrov-)Galerkin schemes the occurrence of such phenomena is prevented by suitable additional assumptions on the operator A, typically from coercivity of the operator [28, 20, 75, 74].

In the entirety of the following discussion, such occurrences as described by Lemma 3.3.4 and discussed in the previous examples are not a-priori excluded. Therefore the vector $f^{(N)}\in\mathbb{C}^N$ must be regarded as an approximate solution to the finite-dimensional truncated problem in the following sense

$$A_N f^{(N)} = g_N + \varepsilon^{(N)}\,,\quad \text{for some } \varepsilon^{(N)}\in\mathbb{C}^N\,. \tag{3.12}$$

As a notational matter, the symbol $\varepsilon^{(N)}$ is used rather than ε_N to emphasise there is no reason whatsoever that the residual vector $\varepsilon^{(N)}$ in the N-dimensional problem is a truncation for every N of the same infinite-dimensional vector $\varepsilon \in \mathcal{H}$. It may well be that $\varepsilon^{(N)} = 0$ for some values of N, namely where a solution $f^{(N)} \in \mathbb{C}^N$ exists to the problem (3.6), whereas for other values of N, $\varepsilon^{(N)} \neq 0$. This is typical of the case where A_N alternates between a non-singular and singular matrix respectively.

It is desirable to assume that $\varepsilon^{(N)}$ is small and asymptotically vanishes with N, or even that $\varepsilon^{(N)} = 0$ for N large enough. The condition that $\varepsilon^{(N)}$ vanishes as $N \to \infty$ is analogous to the assumption of *asymptotic consistency* for the weak formulation Galerkin projection methods in finite-element theory [28, Definition 2.15]. In the adaptation to this work, the term *'asymptotic consistency'* is used to describe the situation in which $\varepsilon^{(N)} \xrightarrow{N\to\infty} 0$. This is motivated by the following lemma.

Lemma 3.3.5. *Let $A \in \mathscr{B}(\mathcal{H})$ and $g \in \mathrm{ran}A$. Let A_N and g_N be defined as in* (3.5) *and* (3.4) *respectively, for $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ orthonormal bases of $\mathcal{H}$. Then there always exists a sequence $(f^{(N)})_{N\in\mathbb{N}}$ such that*

$$f^{(N)} \in \mathbb{C}^N \quad \textit{and} \quad \lim_{N\to\infty} \left\|A_N f^{(N)} - g_N\right\|_{\mathbb{C}^N} = 0\,.$$

In other words, there exist approximate solutions $(f^{(N)})_{N\in\mathbb{N}}$ to (3.12) *that satisfy the assumption of asymptotic consistency, i.e. $\left\|\varepsilon^{(N)}\right\|_{\mathbb{C}^N} \xrightarrow{N\to\infty} 0$.*

Proof. Let f be a solution to $Af = g$. Then the sequence $(f^{(N)})_{N\in\mathbb{N}}$ defined by

$$f^{(N)} := (P_N f)_N = f_N\,,$$

that is, $\widehat{f^{(N)}} = P_N f$, does the trick. To show this claim, it suffices to note that $Q_N A P_N \to A$ in the strong operator topology (see Lemma 3.5.1). Therefore, by adding and subtracting $Af = g$,

$$\begin{aligned} \left\|A_N f^{(N)} - g_N\right\|_{\mathbb{C}^N} &= \left\|Q_N A P_N \widehat{f^{(N)}} - Q_N g\right\|_{\mathcal{H}} \\ &= \|Q_N A P_N f - Af + Af - Q_N g\|_{\mathcal{H}} \\ &\leq \|(Q_N A P_N - A)f\|_{\mathcal{H}} + \|Af - Q_N g\|_{\mathcal{H}} \\ &= \|(Q_N A P_N - A)f\|_{\mathcal{H}} + \|g - Q_N g\|_{\mathcal{H}} \\ &= \|(Q_N A P_N - A)f\|_{\mathcal{H}} + \|(\mathbb{1} - Q_N)g\|_{\mathcal{H}}\,. \end{aligned}$$

Taking $N \to \infty$, the strong limit yields the conclusion. □

3.3.2 Error and residual convergence of the truncated problem

The major question concerning approximate solutions to infinite-dimensional linear inverse problems is the 'goodness' of the approximation by the method. That is to say, whether natural indicators of the difference between the infinite-dimensional inverse linear problem and its finite-dimensional truncation, namely the error (3.8) and residual (3.9), converge in some sense to zero as $N \to \infty$.

The most obvious obstruction to $\mathscr{E}_N$ vanishing in the limit $N \to \infty$, is when the orthonormal system $(u_n)_{n\in\mathbb{N}}$ does *not possess the approximability property*, i.e. $\operatorname{span}\{u_n \,|\, n \in \mathbb{N}\}$ is *not* dense in $\mathcal{H}$. The following simple example illustrates such an issue.

Example 3.3.6. Consider the weighted (compact) right shift operator $\mathcal{R}$ on $\ell^2(\mathbb{N})$ (see Appendix A). If this is truncated with respect to the orthonormal system

$$(u_n)_{n\in\mathbb{N}} = (e_n)_{n\geq 2}\,, \quad (v_n)_{n\in\mathbb{N}} = (e_n)_{n\in\mathbb{N}}$$

and the inverse linear problem is $\mathcal{R}f = g = e_2$, then the exact solution is $f = \frac{1}{\sigma_1}e_1$. But, the truncated problem produces the approximate solutions

$$\widehat{f^{(N)}} \in \operatorname{span}\{e_2, e_3, \dots\}\,,$$

so that $f \perp \widehat{f^{(N)}}$ for all $N \in \mathbb{N}$ and so $\left\|\widehat{f^{(N)}} - f\right\|_{\mathcal{H}} \geq \frac{1}{\sigma_1}$.

Although the above example, and other similar situations where one does not truncate with a complete basis of $\mathcal{H}$, appears unwise, in certain contexts it is natural. For example, within the framework of Krylov subspace projection methods (see Chapters 2 and 4 for more details) it is not a-priori guaranteed that the vectors spanning the search space, here the Krylov subspace, are dense in $\mathcal{H}$. It may well happen that $\overline{\mathcal{K}(A,\, g)} \subsetneq \mathcal{H}$. Recall that, as defined in Chapter 2, the Krylov subspace with respect to the operator A and a vector

g is

$$\mathcal{K}\,(A,\,g) := \operatorname{span}\left\{A^n g \,|\, n \in \mathbb{N}_0\right\}\ .$$

Below are some simple examples revealing when the Kyrlov subspace may or may not be all of $\mathcal{H}$, that is to say, when the orthonormal system $(u_n)_{n\in\mathbb{N}}$ forms a basis or not for $\mathcal{H}$.

Example 3.3.7. (i) For the right-shift operator R on $\ell^2(\mathbb{N})$ (Appendix A) and the vector $g = e_{m+1}$ (given some $m \in \mathbb{N}$) one has $\overline{\mathcal{K}\,(R,\,g)} = \operatorname{span}\{e_1, \ldots, e_m\}^{\perp}$ (c.f. Example 4.4.1 (iv)). Clearly then $\overline{\mathcal{K}\,(R,\,g)}$ is a proper subspace of $\ell^2(\mathbb{N})$ if $m > 1$, and all of $\ell^2(\mathbb{N})$ when $m = 1$. The exact solution to the inverse linear problem $Rf = g$ for $g \in \operatorname{ran}R$ is *not* solvable using projection methods based on Krylov subspaces.

(ii) For the Volterra operator on $L^2[0,1]$ (Appendix A), and the function $g = \mathbf{1}$, it follows that $\overline{\mathcal{K}\,(V,\,g)} = L^2[0,1]$ (see Example 4.2.2 (iii)). A projection method using this subspace as the search and trial space would then possess the approximability property.

Of course, from Definition 3.2.4, the lack of the approximability property is ruled out as both $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ must be orthonormal bases of the space $\mathcal{H}$.

3.4 Compact linear inverse problems

A natural class of inverse problems to investigate with the theoretical framework laid out so far, are the compact linear inverse problems. Compact operators in a separable Hilbert space $\mathcal{H}$ can be approximated in the operator norm by finite rank operators, also known as *degenerate operators* (see Lemma 3.4.1 below). In general Hilbert spaces, the space of compact operators is *closed* in the space of bounded operators equipped with the operator norm topology [51]. Also, the space of degenerate operators is a linear manifold contained within the space of compact operators, but in general it is *not* closed [51].

A compact operator A on a separable Hilbert space $\mathcal{H}$ admits the following canonical decomposition, known as the *singular value decomposition* or *canonical expansion* [51, Chapter V, equation (2.23)]

$$A = \sum_{n\in\mathcal{J}} \sigma_n \left|\psi_n\right\rangle \left\langle\varphi_n\right| \,, \tag{3.13}$$

where n runs over $\mathcal{J} \subset \mathbb{N}$, with $\sup\mathcal{J} < \infty$ or $\sup\mathcal{J} = \infty$, $\sigma_n \geq \sigma_{n+1} > 0$ for all $n \in \mathcal{J}$, $\sigma_n \xrightarrow{n\to\infty} 0$, and $(\psi_n)_n$ and $(\varphi_n)_n$ are two orthonormal systems of $\mathcal{H}$. The series (3.13) converges in operator norm.

Injectivity (respectively dense range in $\mathcal{H}$) of the operator A is equivalent to $(\varphi_n)_{n\in\mathbb{N}}$ (respectively $(\psi_n)_{n\in\mathbb{N}}$) forming an orthonormal basis of $\mathcal{H}$.

Given that $\dim\mathcal{H} = \infty$, the compactness of an injective A ensures that A^{-1} exists only on the range, and may *not* have an everywhere defined, bounded inverse. $\mathrm{ran}A$ may be dense in $\mathcal{H}$, for example the Volterra operator on $L^2[0,1]$ (see Appendix A), but $\mathrm{ran}A$ may never be the entire Hilbert space $\mathcal{H}$. It may also be that $\mathrm{ran}A$ is dense in a closed subspace of the ambient Hilbert space, for example the weighted right-shift operator on $\ell^2(\mathbb{N})$.

The following lemma reveals that the compression of the compact operator A, namely Q_NAP_N (as in (3.3)), is close to the operator A in operator norm.

Lemma 3.4.1. *Let $\mathcal{H}$ be an infinite-dimensional separable Hilbert space, let $A : \mathcal{H} \to \mathcal{H}$ be a compact linear operator, and let $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be two orthonormal bases for $\mathcal{H}$. Then*

$$\|A - Q_NAP_N\|_{\mathrm{op}} \xrightarrow{N\to\infty} 0 \,, \tag{3.14}$$

where Q_N and P_N are the orthogonal projections defined by (3.2).

Proof. Split the term $A - Q_NAP_N$ as follows

$$A - Q_NAP_N = (A - Q_NA) + Q_N(A - AP_N) \,,$$

so that it suffices only to prove that $\|A - Q_NA\|_{\mathrm{op}} \to 0$ and $\|A - AP_N\|_{\mathrm{op}} \to 0$ as $N \to \infty$. For now the focus will remain on the term $A - AP_N$.

To this end, it is trivial to see that any compact operator A in $\mathcal{H}$, as given by (3.13), is arbitrarily well approximated in the operator norm by a finite rank operator $\widetilde{A}$, say

$$\widetilde{A} = \sum_{n=1}^{M} \sigma_n \left|\psi_n\right\rangle \left\langle\varphi_n\right| \,,$$

for some $M \in \mathbb{N}$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_M \geq 0$, and $(\psi_n)_{n=1}^{M}$ and $(\varphi_n)_{n=1}^{M}$ are the appropriate orthonormal systems from (3.13). Therefore, one has

$$\begin{aligned} A - AP_N &= A - \widetilde{A} + \widetilde{A} - \widetilde{A}P_N + \widetilde{A}P_N - AP_N \\ &= (A - \widetilde{A}) + (\widetilde{A} - \widetilde{A}P_N) - (A - \widetilde{A})P_N \,, \end{aligned}$$

and from the triangle inequality one has

$$\begin{aligned} \|A - AP_N\|_{\mathrm{op}} &\leq \left\|A - \widetilde{A}\right\|_{\mathrm{op}} + \left\|\widetilde{A} - \widetilde{A}P_N\right\|_{\mathrm{op}} + \left\|A - \widetilde{A}\right\|_{\mathrm{op}} \|P_N\|_{\mathrm{op}} \\ &\leq 2\left\|A - \widetilde{A}\right\|_{\mathrm{op}} + \left\|\widetilde{A} - \widetilde{A}P_N\right\|_{\mathrm{op}} \,. \end{aligned}$$

As the approximation between $\widetilde{A}$ and A is arbitrarily small, the focus is now to show that $\left\|\widetilde{A} - \widetilde{A}P_N\right\|_{\mathrm{op}}$ vanishes in the limit $N \to \infty$. Taking some generic $\xi = \sum_{k=1}^{\infty} \xi_n u_n \in \mathcal{H}$ such that $\|\xi\|_{\mathcal{H}} = 1$ one has that

$$\begin{aligned} \left\|(\widetilde{A} - \widetilde{A}P_N)\xi\right\|_{\mathcal{H}}^2 &= \left\| \sum_{n=1}^{M} \sigma_n \left(\sum_{k=N+1}^{\infty} \xi_k \left\langle \varphi_n, \, u_k \right\rangle \right) \psi_n \right\|_{\mathcal{H}}^2 \\ &= \sum_{n=1}^{M} \sigma_n^2 \left| \sum_{k=N+1}^{\infty} \xi_k \left\langle \varphi_n, \, u_k \right\rangle \right|^2 \leq \sum_{n=1}^{M} \sigma_n^2 \, \|(\mathbb{1} - P_N)\varphi_n\|_{\mathcal{H}}^2 \,, \end{aligned}$$

and so

$$\left\|\widetilde{A} - \widetilde{A}P_N\right\|_{\mathrm{op}}^2 \leq M\sigma_1^2 \max_{n \in \{1,\ldots,M\}} \|(\mathbb{1} - P_N)\varphi_n\|_{\mathcal{H}}^2 \xrightarrow{N\to\infty} 0 \,,$$

since the maximum is taken over M finitely many quantities, each of which vanishes in the limit as $N \to \infty$.

The vanishing of $\|A - Q_N A\|_{\mathrm{op}}$ is proven in much the same way. Splitting as follows

$$A - Q_N A = (A - \widetilde{A}) + (\mathbb{1} - Q_N)\widetilde{A} - Q_N(A - \widetilde{A})\,,$$

so that $\|A - Q_N A\|_{\mathrm{op}} \leq 2\left\|A - \widetilde{A}\right\|_{\mathrm{op}} + \left\|(\mathbb{1} - Q_N)\widetilde{A}\right\|_{\mathrm{op}}$. Controlling the vanishing of $(\mathbb{1} - Q_N)\widetilde{A}$ in the operator norm immediately gives the required convergence result. Again, suppose $\xi = \sum_{n\in\mathbb{N}} \xi_n \varphi_n \in \mathcal{H}$ is such that $\|\xi\|_{\mathcal{H}} = 1$. As $\widetilde{A}\xi$ is a vector in $\mathcal{H}$ and is independent of N,

$$\left\|(\mathbb{1} - Q_N)\widetilde{A}\xi\right\|_{\mathcal{H}}^2 = \left\|\sum_{k=N+1}^{\infty}\sum_{n=1}^{M} \sigma_n \xi_n \langle \psi_n,\, v_k\rangle\, v_k\right\|_{\mathcal{H}}^2 ,$$

so that as above, eventually one has

$$\begin{aligned}\left\|(\mathbb{1} - Q_N)\widetilde{A}\right\|_{\mathrm{op}}^2 &\leq \sum_{k=N+1}^{\infty}\sum_{n=1}^{M} \sigma_n^2 \left|\langle \psi_n,\, v_k\rangle\right|^2 \\ &\leq M\sigma_1^2 \max_{n\in\{1,\dots,M\}} \|(\mathbb{1} - Q_N)\psi_n\|_{\mathcal{H}}^2 \to 0\,,\end{aligned}$$

as $N \to \infty$. □

An alternative form of the proof of Lemma 3.4.1 is provided by [53, Chapter 4, Lemma 15.4].

The following result describes the *generic* convergence behaviour of the well-defined compact inverse problem under a projection method.

Theorem 3.4.2. *Consider the linear inverse problem $Af = g$ in a separable Hilbert space $\mathcal{H}$ for some compact and injective operator $A : \mathcal{H} \to \mathcal{H}$ and some $g \in \mathrm{ran}A$; and the associated finite-dimensional truncation A_N obtained by compression with respect to the orthonormal bases $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ of $\mathcal{H}$.*

Let $(f^{(N)})_{N\in\mathbb{N}}$ be a sequence of approximate solutions to the truncated problem in the quantitative sense

$$A_N f^{(N)} = g_N + \varepsilon^{(N)} \qquad f^{(N)}, \varepsilon^{(N)} \in \mathbb{C}^N \qquad \left\|\varepsilon^{(N)}\right\|_{\mathbb{C}^N} \xrightarrow{N\to\infty} 0$$

for every (sufficiently large) N. If $\widehat{f^{(N)}}$ is $\mathcal{H}$-norm uniformly bounded in N, then

$$\|\mathfrak{R}_N\|_{\mathcal{H}} \to 0 \quad \textit{and} \quad \mathscr{E}_N \rightharpoonup 0$$

as $N \to \infty$.

Remark 3.4.3. Although the vectors used to construct the compression of the operator form an orthonormal basis of the underlying Hilbert space, the theorem does *not* generally assume that the truncated finite-dimensional problem posed in $\mathbb{C}^N$ is solvable at *every* N, and hence one is no longer in the Petrov-Galerkin scheme, but rather in the scheme of general projection methods.

Proof of Theorem 3.4.2. Split $A\widehat{f^{(N)}} - g$ as follows

$$(*) \qquad A\widehat{f^{(N)}} - g = (A - Q_N A P_N)\widehat{f^{(N)}} + Q_N A P_N \widehat{f^{(N)}} - Q_N g + Q_N g - g\,.$$

By assumption $\|Q_N g - g\|_{\mathcal{H}} \xrightarrow{N\to\infty} 0$ and

$$\begin{aligned}\left\|Q_N A P_N \widehat{f^{(N)}} - Q_N g\right\|_{\mathcal{H}} &= \|A_N f^{(N)} - g_N\|_{\mathbb{C}^N} \\ &= \|\varepsilon^{(N)}\|_{\mathbb{C}^N} \xrightarrow{N\to\infty} 0\,.\end{aligned}$$

Lemma 3.4.1 and the assumption of uniform boundedness with N of the sequence $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$, imply that

$$\left\|(A - Q_N A P_N)\widehat{f^{(N)}}\right\|_{\mathcal{H}} \leq \|A - Q_N A P_N\|_{\mathrm{op}} \left\|\widehat{f^{(N)}}\right\|_{\mathcal{H}} \xrightarrow{N\to\infty} 0\,.$$

Using this information and the triangle inequality in (*), one immediately obtains that $\|\mathfrak{R}_N\|_{\mathcal{H}} \to 0$ as $N \to \infty$.

Now, exploiting the singular value decomposition (3.13) of A, where $(\varphi_n)_{n\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}$ and $(\psi_n)_{n\in\mathbb{N}}$ is a orthonormal system of $\mathcal{H}$, and $0 < \sigma_{n+1} \leq \sigma_n$ $\forall n \in \mathcal{J}$, then

$$\widehat{f^{(N)}} = \sum_{n\in\mathbb{N}} f_n^{(N)} \varphi_n\,, \quad f = \sum_{n\in\mathbb{N}} f_n \varphi_n\,,$$

from which

$$0 = \lim_{N\to\infty} \left\| A\widehat{f^{(N)}} - g \right\|_{\mathcal{H}}^2 = \lim_{N\to\infty} \sum_{n\in\mathcal{J}} \sigma_n^2 \left| f_n^{(N)} - f_n \right|^2 .$$

Then it is obvious that $\widehat{f^{(N)}}$ converges to f component-wise, i.e., $\mathscr{E}_N \rightsquigarrow 0$, as $\mathcal{J} = \mathbb{N}$ due to injectivity.

By assumption, $\widehat{f^{(N)}}$ remains uniformly bounded in $\mathcal{H}$, thus $\widehat{f^{(N)}} \rightharpoonup f$ (Lemma C.1.3). □

Theorem 3.4.2 provides sufficient conditions for some form of vanishing of the error and residual. The key assumptions are: injectivity of A, asymptotic consistency of the truncated problems, and uniform boundedness of the approximate solutions. In fact, injectivity was only used to ensure the convergence of the error term, however the residual converges *regardless* of this assumption. From the iterates $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$ there exists another sequence in the Hilbert space made from convex combinations of the $\widehat{f^{(N)}}$'s that converges *strongly* to the solution f [10, Corollary 3.8].

Theorem 3.4.2 is analogous to a result known as the 'Fundamental Convergence Theorem' ([53, Chapter 4, Theorems 15.1 and 15.2]), which gives necessary and sufficient conditions for the convergence of the residual term for Petrov-Galerkin projection methods.

Remark 3.4.4. Under the conditions of Theorem 3.4.2, the strong vanishing of the residual $\mathfrak{R}_N$ and the weak vanishing of the error $\mathscr{E}_N$ is a *generic* behaviour. For example, the compact inverse problem $\mathcal{R}f = 0$ in $\ell^2(\mathbb{N})$ associated with the weighted right-shift $\mathcal{R}$, with $\sigma_n \neq 0$ for all n, has exact solution $f = 0$. The truncated problem $\mathcal{R}_N f^{(N)} = 0$ with respect to the same basis $(e_n)_{n\in\mathbb{N}}$, $\mathcal{R}_N$ being the matrix in (3.10), is solved by $\mathbb{C}^N$-vectors whose first $N-1$ components are zero, $\widehat{f^{(N)}} = e_N$. The sequence $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$ converges weakly to zero in $\ell^2(\mathbb{N})$, so indeed $\mathscr{E}_N \rightharpoonup 0$, and also by compactness $\mathfrak{R}_N \to 0$. But $\|\mathscr{E}_N\|_{\mathcal{H}} = 1$ for all N, so the error cannot vanish in the strong topology.

Remark 3.4.5. From the example in Remark 3.4.4 one may see how it may happen that the solutions selected are 'bad' approximate solutions so that $\|f^{(N)}\|_{\mathbb{C}^N} = \left\|\widehat{f^{(N)}}\right\|_{\mathcal{H}} \to \infty$, even though the 'good' property $\|A_N f^{(N)} - g_N\|_{\mathbb{C}^N}$ vanishing to 0 is preserved. For instance, if one should choose the solution $\widehat{f^{(N)}} = \sigma_N^{-\frac{1}{2}} e_N$. From compactness of $\mathcal{R}$ it is known that $\sigma_N \to 0$, and one has that $\mathcal{R}_N f^{(N)} = 0$ along with $\mathcal{R}\widehat{f^{(N)}} = \sigma_N^{\frac{1}{2}} e_N$ so that $\mathfrak{R}_N \to 0$. Yet, $\left\|\widehat{f^{(N)}}\right\|_{\mathcal{H}} \to \infty$. Therefore, the assumption of uniform boundedness of the approximate solutions in Theorem 3.4.2 is not redundant. This small example also shows that, although by compactness $\widehat{f^{(N)}} \rightharpoonup f$ implies $\left\|A\widehat{f^{(N)}} - Af\right\|_{\mathcal{H}} \to 0$, the opposite implication is generally false.

Remark 3.4.6. Even if the genericity in Remarks 3.4.4 and 3.4.5 is referred to compact injective operators without dense range, requiring $\overline{\mathrm{ran} A} = \mathcal{H}$ does not improve the convergence in general. For example, the compact inverse problem associated to the right-shift $\mathcal{R}$ in $\ell^2(\mathbb{Z})$ (Appendix A) involves a compact, injective operator with dense range. But, again the compression with $Q_N := P_N := \sum_{n=-N}^{n=N} |e_n\rangle\langle e_n|$ produces for every $N \in \mathbb{N}$ a $(2N+1)\times(2N+1)$ square matrix that is singular, and therefore the considerations of Remarks 3.4.4 and 3.4.5 may be repeated verbatim.

Remark 3.4.7. In Lemma 3.3.4 it is shown that 'bad' truncation schemes are always possible, i.e., truncations that lead to matrices A_N that are, eventually in N, always singular. On the other hand, there always exists a 'good' choice for the truncation which makes the infinite-dimensional residual and error vanish in a stronger sense that given in Theorem 3.4.2 (although this choice may not always be explicitly identifiable). Moreover, this choice of truncation scheme does not require the extra assumption of the uniform boundedness of the approximate solutions $\widehat{f^{(N)}}$. For instance, in terms of the singular value decomposition (3.13) of A, it suffices to choose

$$(u_n)_{n\in\mathbb{N}} = (\varphi_n)_{n\in\mathbb{N}} \quad (v_n)_{n\in\mathbb{N}} = (\psi_n)_{n\in\mathbb{N}} \,,$$

so that $Q_N A P_N = \sum_{n=1}^{N} \sigma_n |\psi_n\rangle\langle\varphi_n|$ and $A_N = \mathrm{diag}(\sigma_1, \dots, \sigma_N)$. So for given $g = \sum_{n\in\mathbb{N}} g_n \psi_n \in \mathrm{ran}A$, one has $\widehat{f^{(N)}} = \sum_{n=1}^{N} \frac{g_n}{\sigma_n}\varphi_n$, where the sequence

$\left(\frac{g_n}{\sigma_n}\right)_{n\in\mathbb{N}} \in \ell^2(\mathbb{N})$ as $g \in \mathrm{ran}A$. So

$$\left\|f - \widehat{f^{(N)}}\right\|_{\mathcal{H}}^2 = \sum_{n=N+1}^{\infty} \left|\frac{g_n}{\sigma_n}\right|^2 \xrightarrow{N\to\infty} 0\,.$$

In addition to Remark 3.4.7, a particular projection method that does result in a good truncation scheme for the compact inverse problem is the conjugate-gradient method. The strong vanishing of the error term is guaranteed under the assumption that $g \in \mathrm{ran}A$ (see Theorem 2.3.6 in Chapter 2).

3.5 Bounded linear inverse problems

This Section serves as a comparison between the findings for the compact inverse problem and the more general case of the bounded linear inverse problem.

The first result presented here is that, unlike the case for compact linear operators, the convergence of the compression Q_NAP_N to the operator A is no longer controlled in the operator norm when $\dim\mathcal{H} = \infty$, but is controlled in the strong operator topology. That is to say, given a vector in $\psi \in \mathcal{H}$, the $\mathcal{H}$-norm vanishing of $(Q_NAP_N - A)\psi$ occurs.

Lemma 3.5.1. *Let $\mathcal{H}$ be a separable Hilbert space, and let $A \in \mathscr{B}(\mathcal{H})$. Let $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ be orthonormal bases of $\mathcal{H}$, and define the orthogonal projection operators Q_N and P_N as in* (3.2). *Then $Q_NAP_N \to A$ as $N \to \infty$ in the strong operator topology, i.e., given some $\psi \in \mathcal{H}$ one has $\|(Q_NAP_N - A)\psi\|_{\mathcal{H}} \xrightarrow{N\to\infty} 0$.*

Proof. Consider that one may write

$$\begin{aligned}(Q_NAP_N - A)\psi &= (Q_NA - A)\psi + (Q_NAP_N - Q_NA)\psi \\ &= (Q_N - \mathbb{1})A\psi + Q_NA(P_N - \mathbb{1})\psi\,,\end{aligned}$$

so that $\|(Q_NAP_N - A)\psi\|_{\mathcal{H}} \leq \|(Q_N - \mathbb{1})A\psi\|_{\mathcal{H}} + \|A\|_{\mathrm{op}}\,\|(P_N - \mathbb{1})\psi\|_{\mathcal{H}}$ for any $\psi \in \mathcal{H}$. Clearly then $\|(Q_NAP_N - A)\psi\|_{\mathcal{H}} \xrightarrow{N\to\infty} 0$. □

The lack of operator norm convergence is obvious by considering, for instance, the compression of the identity operator on $\mathcal{H}$. The operator norm limit of degenerate (finite-rank) operators can only be compact.

Therefore, the control of the infinite-dimensional linear inverse problem in terms of finite dimensional truncations is, in general, less strong than the compact inverse problem counterpart.

The following theorem presents a counterpart result to Theorem 3.4.2 for the generic behaviour of well-defined bounded inverse problems.

Theorem 3.5.2. *Consider the linear inverse problem $Af = g$ in a separable Hilbert space $\mathcal{H}$ for some bounded and injective $A : \mathcal{H} \to \mathcal{H}$ and some $g \in \mathcal{H}$, and the finite-dimensional truncation A_N obtained by compression with respect to the orthonormal bases $(u_n)_{n\in\mathbb{N}}$ and $(v_n)_{n\in\mathbb{N}}$ of $\mathcal{H}$.*

Let $(f^{(N)})_{N\in\mathbb{N}}$ be a sequence of approximate solutions to the truncated problems in the quantitative sense

$$A_N f^{(N)} = g_N + \varepsilon^{(N)}\,, \quad f^{(N)}, \varepsilon^{(N)} \in \mathbb{C}^N\,, \quad \|\varepsilon^{(N)}\|_{\mathbb{C}^N} \xrightarrow{N\to\infty} 0$$

for every (sufficiently large) N. Assume further that $\widehat{f^{(N)}}$ converges strongly in $\mathcal{H}$, equivalently, that $\|f^{(N)} - f^{(M)}\|_{\mathbb{C}^{\max\{N,M\}}} \xrightarrow{N,M\to\infty} 0$. Then

$$\|\mathscr{E}_N\|_{\mathcal{H}} \to 0 \quad \textit{and} \quad \|\mathfrak{R}_N\|_{\mathcal{H}} \to 0 \quad \textit{as } N \to \infty\,.$$

Proof. Splitting $A\widehat{f^{(N)}} - g$ as follows

$$\begin{aligned} A\widehat{f^{(N)}} - g = &(A - Q_N A P_N)\widehat{f^{(N)}} \\ &+ Q_N A P_N \widehat{f^{(N)}} - Q_N g \\ &+ Q_N g - g\,, \end{aligned} \tag{**}$$

and by assumption $\|Q_N g - g\|_{\mathcal{H}} \xrightarrow{N\to\infty} 0$ and

$$\begin{aligned} \Big\| Q_N A P_N \widehat{f^{(N)}} - Q_N g \Big\|_{\mathcal{H}} &= \|A_N f^{(N)} - g_N\|_{\mathbb{C}^N} \\ &= \|\varepsilon^{(N)}\|_{\mathbb{C}^N} \xrightarrow{N\to\infty} 0\,, \end{aligned}$$

so the strong vanishing of $A\widehat{f^{(N)}} - g$ is the same as the strong vanishing of $(A - Q_N A P_N)\widehat{f^{(N)}}$.

By assumption, $\left\|\widehat{f^{(N)}} - \widetilde{f}\right\|_{\mathcal{H}} \xrightarrow{N\to\infty} 0$ for some $\widetilde{f} \in \mathcal{H}$, so

$$\begin{aligned}\left\|(A - Q_N A P_N)\widehat{f^{(N)}}\right\|_{\mathcal{H}} &= \left\|(A - Q_N A P_N)\widetilde{f} + (A - Q_N A P_N)(\widehat{f^{(N)}} - \widetilde{f})\right\|_{\mathcal{H}} \\ &\leq \left\|(A - Q_N A P_N)\widetilde{f}\right\|_{\mathcal{H}} \\ &\qquad + \|A - Q_N A P_N\|_{\text{op}} \left\|\widehat{f^{(N)}} - \widetilde{f}\right\|_{\mathcal{H}} \\ &\leq \left\|(A - Q_N A P_N)\widetilde{f}\right\|_{\mathcal{H}} + 2\,\|A\|_{\text{op}} \left\|\widehat{f^{(N)}} - \widetilde{f}\right\|_{\mathcal{H}}\end{aligned}$$

so that $\left\|(A - Q_N A P_N)\widehat{f^{(N)}}\right\|_{\mathcal{H}} \xrightarrow{N\to\infty} 0$ from Lemma 3.5.1 combined with the strong convergence of $\widehat{f^{(N)}}$. Therefore, (**) implies that $\|\mathfrak{R}_N\|_{\mathcal{H}} \to 0$ as $N \to \infty$.

Furthermore, as $A\widehat{f^{(N)}} \to g$, and $A\widehat{f^{(N)}} \to A\widetilde{f}$ due to continuity of A, so $A\widetilde{f} = g = Af$ and by injectivity one has $f = \widetilde{f}$. Therefore $\|\mathscr{E}_N\|_{\mathcal{H}} \to 0$ as $N \to \infty$. □

In the proof of Theorem 3.5.2, injectivity is only used to show that the error term strongly vanishes, but this information is not needed to prove the strong vanishing of the residual.

In comparing Theorem 3.4.2 with Theorem 3.5.2, injectivity and asymptotic solvability of the truncated problems are *common assumptions* to these theorems. Injectivity merely ensures the solution to $Af = g$ is unique, and yet the asymptotic solvability of the truncated problem is quite a natural assumption too, by virtue of Lemma 3.3.5. But, when passing from the compact case to the general bounded case, one must strengthen the assumption on the $\widehat{f^{(N)}}$'s, namely from being uniformly bounded in N (compact inverse problem) to being strongly convergent (generic bounded inverse problem), to ensure strong vanishing of the residual. As a by-product of the strong convergence of the $\widehat{f^{(N)}}$'s in Theorem 3.5.2, the error term also vanishes strongly.

The proof of Theorem 3.5.2 elucidates the notion that, when A is injective and the truncated problems are asymptotically solvable, the strong, weak or component-wise vanishing occurs if and only if so too does $(A-Q_NAP_N)\widehat{f^{(N)}}$. Yet in the compact case, $A-Q_NAP_N \to \mathbb{O}$ in operator norm (Lemma 3.4.1) so that it suffices that the $\widehat{f^{(N)}}$'s are uniformly bounded (or at least have increasing norm $\|\widehat{f^{(N)}}\|_{\mathcal{H}}$ compensated by the vanishing of $\|A-Q_NAP_N\|_{\mathrm{op}}$) to ensure that $\|\mathfrak{R}_N\|_{\mathcal{H}} \to 0$. The general bounded case though is controlled by the vanishing of $\|(A-Q_NAP_N)\widehat{f^{(N)}}\|_{\mathcal{H}}$ by the additional requirement that the $\widehat{f^{(N)}}$'s converge strongly.

Should strong convergence of the $\widehat{f^{(N)}}$'s *not* occur, then one should expect that the residual converges only in some weaker sense, which also prevents the error from strong vanishing (otherwise $\|\mathscr{E}_N\|_{\mathcal{H}} \to 0$ would imply $\|\mathfrak{R}_N\|_{\mathcal{H}} \to 0$). The following example now reveals a possibility where one only has weak vanishing of the residual term.

Example 3.5.3. Consider the right-shift operator R on $\ell^2(\mathbb{N})$ (Appendix A). This is an injective operator, and the inverse problem $Rf = g = 0$ admits the unique solution $f = 0$. The truncated finite-dimensional problem induced by the bases $(u_n)_{n\in\mathbb{N}} = (v_n)_{n\in\mathbb{N}} = (e_n)_{n\in\mathbb{N}}$ where $(e_n)_{n\in\mathbb{N}}$ is the canonical basis of $\ell^2(\mathbb{N})$, is governed by the subdiagonal matrix

$$R_N = \begin{pmatrix} 0 & \cdots & \cdots & \cdots & 0 \\ 1 & 0 & \cdots & \cdots & 0 \\ 0 & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \ddots & 0 \\ 0 & 0 & \cdots & 1 & 0 \end{pmatrix}.$$

If one were to consider the sequence of approximate solutions $(\widehat{f^{(N)}})_{N\in\mathbb{N}}$ with $\widehat{f^{(N)}} := e_N$ for each N, then

(i) $R_N f^{(N)} = 0 = g_N$ so the truncated problems are solved exactly,

(ii) $\widehat{f^{(N)}} \rightharpoonup 0$ (weak, not strong, convergence of the approximations),

(iii) $\mathfrak{R}_N = g - R\widehat{f^{(N)}} = -e_{N+1} \rightharpoonup 0$ (weak, not strong, vanishing of the residual).

3.6 Some remarks on linear inverse problems with noise

In many typical applications the linear inverse problem is often plagued by noise, or some sort of error, in the datum g. This Section contains some collected remarks on this topic, within the framework of the theory in the previous sections.

Within the modelling framework of a phenomenon, the linear inverse problem $Af = g$ is well-defined (possibly well-posed), and therefore there is a unique 'input' f for a given 'output' g, with an explicitly known mapping $f \overset{A}{\mapsto} g$. But the knowledge of g obtained from measurement is uncertain, and polluted by some noise.

Consequently, $Af = g$ cannot be studied directly, and one often deals with the possibly ill-defined problem

$$A\widetilde{f} = \widetilde{g}\,, \tag{3.15}$$

where $A : \mathcal{H} \to \mathcal{H}$ is a bounded linear operator on Hilbert space, $\widetilde{f}$ is an unknown for the given (measured) 'noisy' term $\widetilde{g} := g + \nu \in \mathcal{H}$, where the 'noise' vector $\nu \in \mathcal{H}$ is typically small, but is not known a-priori (although a bound on $\|\nu\|_{\mathcal{H}}$ may be known).

If ν (and g) belongs to $\mathrm{ran}A$, so too does $\widetilde{g}$, and there will exist an actual solution $\widetilde{f}$ to (3.15). Immediately, Theorems 3.4.2 and 3.5.2 are applicable to this setting (replacing g with $g + \nu$) and analogously one may speak of an approximate solution $f^{(N)} \in \mathbb{C}^N$ such that

$$A_N f^{(N)} = g_N + \nu_N + \varepsilon^{(N)}\,, \quad \|\varepsilon^{(N)}\|_{\mathbb{C}^N} \xrightarrow{N\to\infty} 0\,. \tag{3.16}$$

Within this framework, Theorems 3.4.2 and 3.5.2 are able to produce a control on the convergence of the "residual plus noise" $(g+\nu) - A\widehat{f^{(N)}}$ and the "error plus noise" $\widetilde{f} - \widehat{f^{(N)}}$ terms. But, this framework only determines convergence properties to the "solution plus noise" term $\widetilde{f}$ and *not the exact solution* f. Yet, this can still be informative should ν be sufficiently small. For example, if $A \in \mathscr{B}(\mathcal{H})$ and $A^{-1} \in \mathscr{B}(\mathcal{H})$, then $\widetilde{f} = A^{-1}(g+\nu)$, from which $\|\widetilde{f} - f\|_{\mathcal{H}} \leq \|A^{-1}\|_{\mathrm{op}} \|\nu\|_{\mathcal{H}}$, so that smallness of $\|\nu\|_{\mathcal{H}}$ in terms of $\|A^{-1}\|_{\mathrm{op}}$ ensures $\widetilde{f}$ is a good estimate of f.

On the other hand, when $\nu \notin \mathrm{ran}A$ and $g \in \mathrm{ran}A$, the problem (3.15) loses solvability, i.e., *there is no exact solution to* (3.15), and one may only have an approximate solution $\widetilde{f}$ satisfying $A\widetilde{f} \approx \widetilde{g}$ (and also $A\widetilde{f} \approx g$ as ν is small).

Following this, some comments will be made on the behaviour of the residual and error associated with f, $\widehat{f^{(N)}}$, g for the case of a *compact and injective* operator $A : \mathcal{H} \to \mathcal{H}$, with $g \in \mathrm{ran}A$.

3.6.1 Typical residual behaviour with noise for compact inverse problems

When the truncated linear inverse problem is solved in the approximate sense by (3.16) and the $\widehat{f^{(N)}}$'s are uniformly bounded in $\mathcal{H}$, then one has that

$$\|\mathfrak{R}_N\|_{\mathcal{H}} = \left\|A\widehat{f^{(N)}} - g\right\|_{\mathcal{H}} \xrightarrow{N\to\infty} \|\nu\|_{\mathcal{H}} \,. \tag{3.17}$$

This is seen by splitting $\mathfrak{R}_N$ as follows

$$\mathfrak{R}_N = (Q_N A P_N - A)\widehat{f^{(N)}} + (Q_N g - Q_N A P_N \widehat{f^{(N)}}) + (g - Q_N g)\,,$$

and noting that $\left\|(Q_N A P_N - A)\widehat{f^{(N)}}\right\|_{\mathcal{H}} \leq \|Q_N A P_N - A\|_{\mathrm{op}} \left\|\widehat{f^{(N)}}\right\|_{\mathcal{H}} \to 0$ (Lemma 3.4.1), $\|g - Q_N g\|_{\mathcal{H}} \to 0$, and

$$\begin{aligned}\left\|Q_N g - Q_N A P_N \widehat{f^{(N)}} - \nu\right\|_{\mathcal{H}} &\leq \left\|A_N f^{(N)} - g_N - \nu_N\right\|_{\mathbb{C}^N} + \|(\mathbb{1} - Q_N)\nu\|_{\mathcal{H}} \\ &= \left\|\varepsilon^{(N)}\right\|_{\mathbb{C}^N} + \|(\mathbb{1} - Q_N)\nu\|_{\mathcal{H}} \to 0\,,\end{aligned}$$

as $N \to \infty$, and therefore one has

$$\|\mathfrak{R}_N - \nu\|_{\mathcal{H}} \to 0 \tag{3.18}$$

as $N \to \infty$. So, *the residual vanishes up to the noise threshold.*

3.6.2 Typical error behaviour with noise for compact inverse problems

In the presence of noise one can no longer expect that $\widehat{f^{(N)}}$ converges to f even component-wise. In particular, the possibility $\mathscr{E}_N \to 0$ or $\mathscr{E}_N \rightharpoonup 0$ immediately violates (3.18).

Therefore, $\|\mathscr{E}_N\|_{\mathcal{H}}$ stays strictly above zero uniformly in N. The *typical* (but not general) behaviour of $\|\mathscr{E}_N\|_{\mathcal{H}}$ is that it *initially decreases for N not too large, reaches a minimum, and then increases (possibly blowing up) for larger N.* This phenomenon is typically known as the *'semi-convergence'* of the error. The behaviour is in contrast to that of $\|\mathfrak{R}_N\|_{\mathcal{H}}$ which typically monotonically decreases to the noise threshold. Clearly then, the value $N_0 \in \mathbb{N}$ when $\|\mathscr{E}_N\|_{\mathcal{H}}$ attains its minimum value provides the best approximant of f in $\mathcal{H}$, namely $\widehat{f^{(N_0)}}$.

For concreteness, the Petrov-Galerkin projection method giving (3.16) (where $\varepsilon^{(N)} = 0$ for all N) is performed with the same bases of the canonical singular value decomposition of A (3.13), namely $(\varphi_n)_{n\in\mathbb{N}}$ and $(\psi_n)_{n\in\mathbb{N}}$). The condition $\nu \in \mathrm{ran}A$ is also assumed (where the generalisation to the condition $\nu \notin \mathrm{ran}A$ is straightforward). These simplifications guarantee that for all N, the matrix $A_N = \mathrm{diag}(\sigma_1, \dots, \sigma_N)$ is non-singular on $\mathbb{C}^N$, as now $Q_N A P_N = \sum_{n=1}^{N} \sigma_n |\psi_n\rangle \langle\varphi_n|$, and (3.16) is solved exactly by

$$\widehat{f^{(N)}} = \sum_{n=1}^{N} \frac{g_n + \nu_n}{\sigma_n} \varphi_n \,,$$

where

$$\nu = \sum_{n\in\mathbb{N}} \nu_n \psi_n \,, \quad g = \sum_{n\in\mathbb{N}} g_n \psi_n \,, \quad f = \sum_{n\in\mathbb{N}} f_n \varphi_n \,, \quad g_n = \sigma_n f_n \,.$$

So now $A_N f^{(N)} = g_N + \nu_N$ (where $\varepsilon^{(N)} = 0$ for all N owing to the use of a Petrov-Galerkin method). Straightforward computations of the residual and error yield

$$\|\mathfrak{R}_N\|_{\mathcal{H}}^2 = \left\|g - A\widehat{f^{(N)}}\right\|_{\mathcal{H}}^2 = \sum_{n=1}^{N} |\nu_n|^2 + \sum_{n=N+1}^{\infty} |g_n|^2 \xrightarrow{N\to\infty} \|\nu\|_{\mathcal{H}}^2 \,,$$

$$\|\mathscr{E}_N\|_{\mathcal{H}}^2 = \left\|f - \widehat{f^{(N)}}\right\|_{\mathcal{H}}^2 = \sum_{n=1}^{N} \frac{|\nu_n|^2}{\sigma_n^2} + \sum_{n=N+1}^{\infty} |f_n|^2 = \alpha(N) + \beta(N)\,,$$

where $\alpha(N) := \sum_{n=1}^{N} \frac{|\nu_n|^2}{\sigma_n^2}$ and $\beta(N) := \sum_{n=N+1}^{\infty} |f_n|^2$.

It is clear that $\beta(N)$ monotonically decreases to zero as $N \to \infty$, but $\alpha(N)$ is monotonically increasing with N. This competing behaviour can produce the phenomenon of semi-convergence in $\|\mathscr{E}_N\|_{\mathcal{H}}$. For example, when f is mainly supported on low modes φ_n's and ν has a long tail on high modes ψ_n's, the initial decrease is observed as $\alpha(N)$ does not change much, however $\beta(N)$ decreases substantially. When N is larger, $\alpha(N)$ increases, while $\beta(N)$ remains more or less around zero. Having assumed that $\nu \in \mathrm{ran}A$, necessarily $\alpha(N) \to \|A^{-1}\nu\|_{\mathcal{H}}^2$ and thus the error term does not explode. If, on the other hand, $\nu \notin \mathrm{ran}A$, one would then have that the series defining $\alpha(N)$ diverges. This behaviour is illustrated with a simple example.

Example 3.6.1. For all $n \in \mathbb{N}$, take

$$\sigma_n = n^{-1}\,, \quad g_n = n^{-2}\,, \quad \nu_n = n^{-\frac{3}{2}}\,.$$

So A is an injective operator, $\|\nu\|_{\mathcal{H}}^2 = \zeta(3) \simeq 1.20$ (where $\zeta(x)$ denotes the Riemann zeta function), and $\nu \notin \mathrm{ran}A$. Then $f_n = n^{-1}$, $\|f\|_{\mathcal{H}}^2 = \beta(0) = \frac{\pi^2}{6}$, and

$$\beta(N) \leq (N+1)^{-2} \to 0\,, \quad \alpha(N) \sim \log(N) \to \infty\,.$$

The typical behaviour is displayed in Figure 3.1.

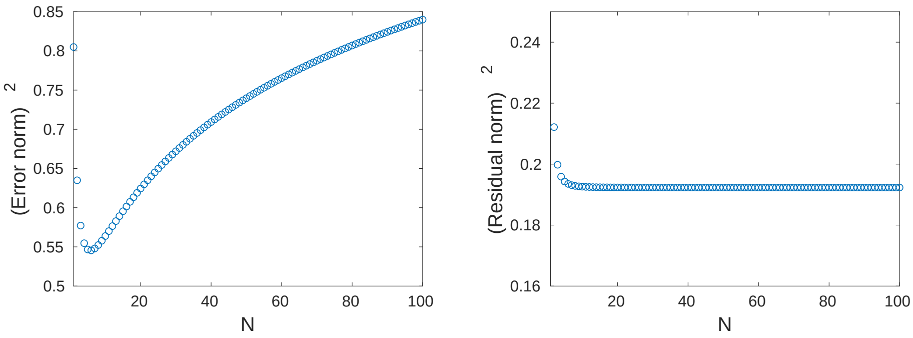


Figure 3.1: Typical behaviour of the error $\|\mathscr{E}_N\|^2_{\mathcal{H}}$ (left) and the residual $\|\mathfrak{R}_N\|^2_{\mathcal{H}}$ (right) with the increasing size of the finite-dimensional truncation, for the problem $Af = g$ considered in Example 3.6.1, and the choice $\sigma_n = n^{-1}$, $g_n = n^{-2}$, and $\nu_n = 0.4n^{-3/2}$.

3.7 Numerical tests & effects of changing truncation basis

Some of the features discussed theoretically in this Chapter will be examined through a few numerical tests concerning different choices of the truncation bases. The bases are Legendre, complex Fourier, and a Krylov basis, that are used to truncate the test problems.

The two model operators considered are the Volterra operator V in $L^2[0,1]$ (Appendix A) and the self-adjoint multiplication operator $M : L^2[1,2] \to L^2[1,2]$, $\psi \mapsto x\psi$. The following two linear inverse problems were examined, namely

1. $Vf_1 = g_1$, with $g_1(x) = \frac{1}{2}x^2$.

   The problem has unique solution

   $$f_1(x) = x\,, \qquad \|f_1\|_{L^2[0,1]} = \frac{1}{\sqrt{3}} \simeq 0.5774 \tag{3.19}$$

   and f_1 *is* a Krylov solution, i.e., $f_1 \in \overline{\mathcal{K}(V,\, g)}$, although $f_1 \notin \mathcal{K}(V,\, g)$ (see Chapter 4, Example 4.4.1 (vii)).

2. $Mf_2 = g_2$, with $g_2(x) = x^2$.

The problem has unique solution

$$f_2(x) = x\,, \quad \|f_2\|_{L^2[1,2]} = \sqrt{\frac{7}{3}} \simeq 1.5275 \tag{3.20}$$

and f_2 *is* a Krylov solution. Indeed,

$$\mathcal{K}\,(M,\,g) = \{x^2 p \,|\, p \in \mathbb{P}_{[1,2]}[x]\}$$

and $\overline{\mathcal{K}\,(M,\,g)} = \{x^2 h(x) \,|\, h \in L^2[1,2]\} = L^2[1,2]$, due to the density of the polynomials in $L^2[1,2]$, from which $f_2 \in \overline{\mathcal{K}\,(M,\,g)}$ and $f_2 \notin \mathcal{K}\,(M,\,g)$.

Both problems were treated with three different orthonormal bases: the Legendre polynomials and the complex Fourier modes (on the intervals $[0,1]$ or $[1,2]$, depending on the problem) solved using the QR factorisation algorithm, and the Krylov basis generated using the GMRES algorithm (with g as the generating vector for the Krylov space).

Computationally speaking, generating accurate representations of the Legendre polynomials is very demanding, and accuracy can be lost rather soon due to their high oscillatory nature (particularly at end points). For this reason, the investigation has been limited to $N = 100$ when considering the Legendre basis, but $N = 500$ when considering the complex Fourier basis. It is expected that there is no significant numerical error from the computation of the Legendre basis, as the $L^2[0,1]$ and $L^2[1,2]$ norms of the basis polynomials have less than 1% error compared to their exact unit value.

For each problem and each choice of basis, monitoring of the norm of the infinite-dimensional error $\|\mathscr{E}_N\|_{L^2} = \|f - \widehat{f^{(N)}}\|_{L^2}$ (for $f = f_1$ or f_2) and the infinite-dimensional residual $\|\mathfrak{R}_N\|_{L^2} = \|g - A\widehat{f^{(N)}}\|_{L^2}$ (for $g = g_1$ or g_2; $A = V$ or M), and the approximated solution $\|\widehat{f^{(N)}}\|_{L^2} = \big\|f^{(N)}\big\|_{\mathbb{C}^N}$ was performed.

Figures 3.2 and 3.4 highlight the difference between the computation in the three bases for the Volterra operator.

(i) In the Legendre basis, $\|\mathscr{E}_N\|_{L^2}$ and $\|\mathfrak{R}_N\|_{L^2}$ are almost zero. $\|\widehat{f^{(N)}}\|_{L^2}$ stays bounded and constant with N and matches the expected value

(3.19).

(ii) In the complex Fourier basis, both $\|\mathscr{E}_N\|_{L^2}$ and $\|\mathfrak{R}_N\|_{L^2}$ are some orders of magnitude *larger* than in the Legendre basis and decrease monotonically with N. In fact, $\|\mathscr{E}_N\|_{L^2}$ and $\|\mathfrak{R}_N\|_{L^2}$ display an evident convergence to zero, however attaining values that are more than ten orders of magnitude larger than the corresponding error and residual norms for the same N is the Legendre case. $\|\widehat{f^{(N)}}\|_{L^2}$, on the other hand, increases monotonically and appears to approach the theoretical value (3.19). These quite stringent differences in the error and residual may be attributable to the Gibbs phenomenon. In fact, reconstructing f_1 using the complex Fourier approximated solutions produces a vector that shows highly oscillatory behaviour near the end points, confirming the presence of Gibbs phenomenon.

(iii) In the Krylov basis $\|\mathscr{E}_N\|_{L^2}$ and $\|\mathfrak{R}_N\|_{L^2}$ decrease monotonically, relatively fast for small N's, then rather slowly with N. These quantaties are smaller than in the complex Fourier basis. $\|\widehat{f^{(N)}}\|_{L^2}$ displays some initial highly oscillatory behaviour, but quickly approaches the theoretical value (3.19). On the other hand, the reconstruction appears to be quite good with some noticeable oscillations near the end points.

Thus, among the considered truncations the Legendre basis yields the most accurate reconstruction and the complex Fourier basis yields the least accurate reconstruction of the exact solution.

In contrast, Figures 3.3 and 3.5 highlight the difference between the computation in the three bases for the M-multiplication operator.

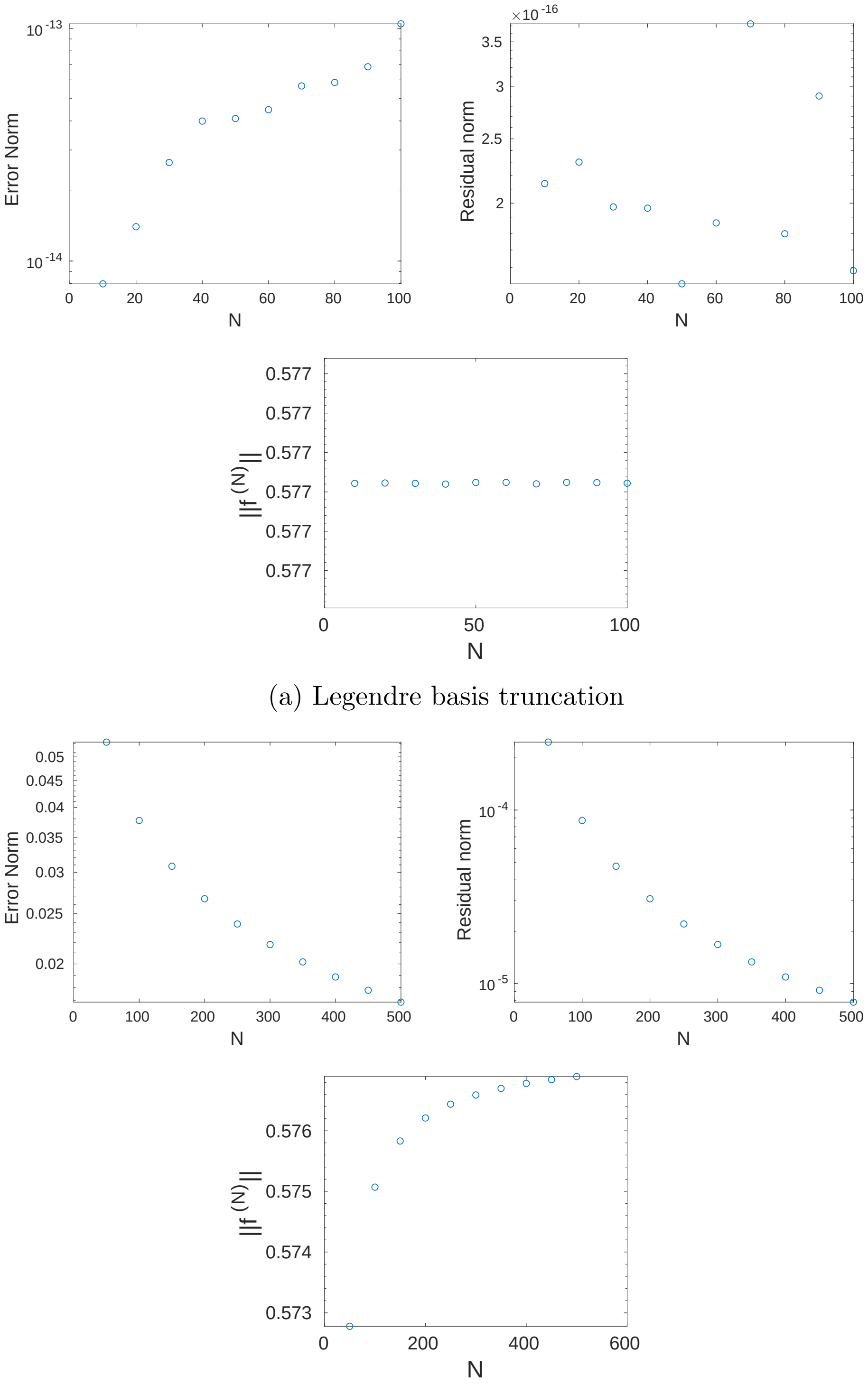
(i) In the Legendre basis, $\|\mathscr{E}_N\|_{L^2}$ and $\|\mathfrak{R}_N\|_{L^2}$ are again almost zero. $\|\widehat{f^{(N)}}\|_{L^2}$ is constant with N at the expected value (3.20). The approximated solutions reconstruct the exact solution f_2 at any truncation number.

(ii) In the Fourier basis the behaviour of the above indicators is again qualitatively the same, again with a much milder convergence rate in N

to the asymptotic values as compared with the Legendre case. $\|\mathscr{E}_N\|_{L^2}$ and $\|\mathfrak{R}_N\|_{L^2}$ still display an evident convergence to zero. Again the higher error compared to the Legendre case is likely due to the nature of the approximation of the exact solution f_2 by oscillatory functions and the Gibbs phenomenon.

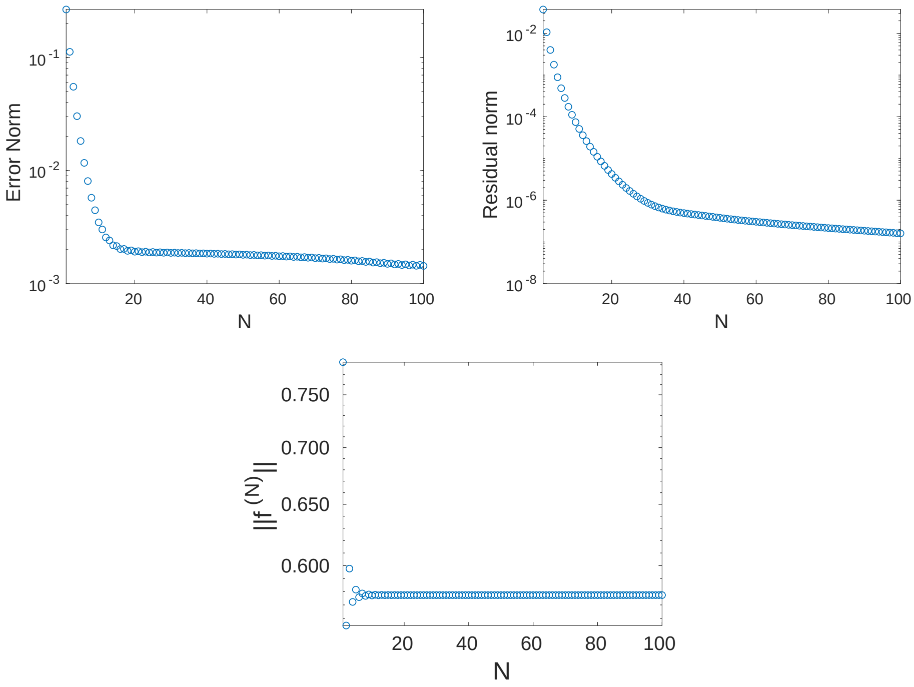
(iii) The Krylov basis displays a fast initial decrease of both $\|\mathscr{E}_N\|_{L^2}$ and $\|\mathfrak{R}_N\|_{L^2}$ to the tolerance level of 10^{-10} that was set for the residual. Also the magnitude of $\|\widehat{f^{(N)}}\|_{L^2}$ increases rapidly and remains constant at the expected value (3.20). The reconstruction of the solution is excellent, but still not quite as good as the Legendre case.

All this gives numerical evidence that the choice of the truncation basis *does* affect the sequence of solutions. The Legendre basis is best suited to these problems as f_1, f_2, g_1 and g_2 are perfectly representable by the first few basis vectors.



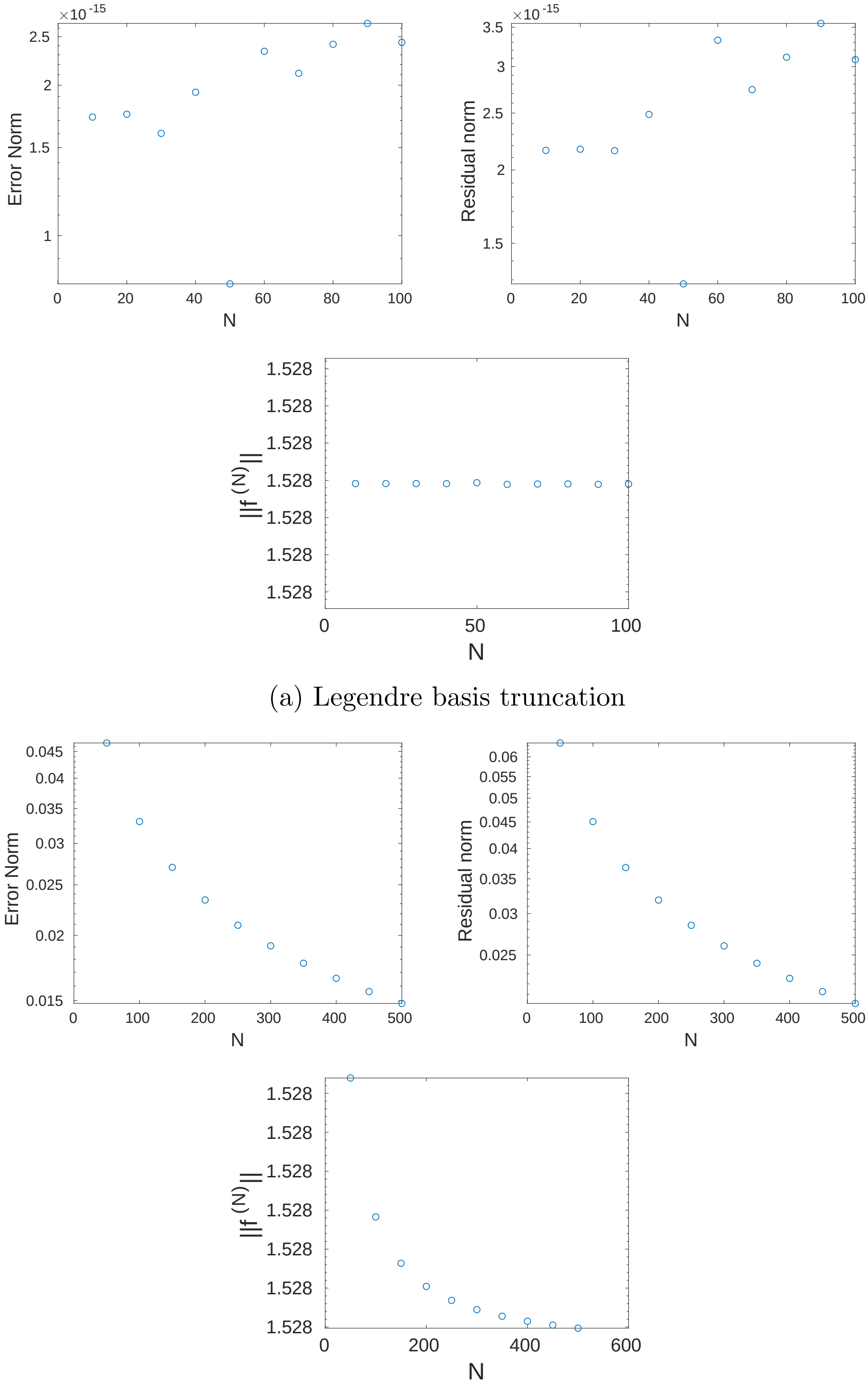
(a) Legendre basis truncation

(b) Complex Fourier basis truncation



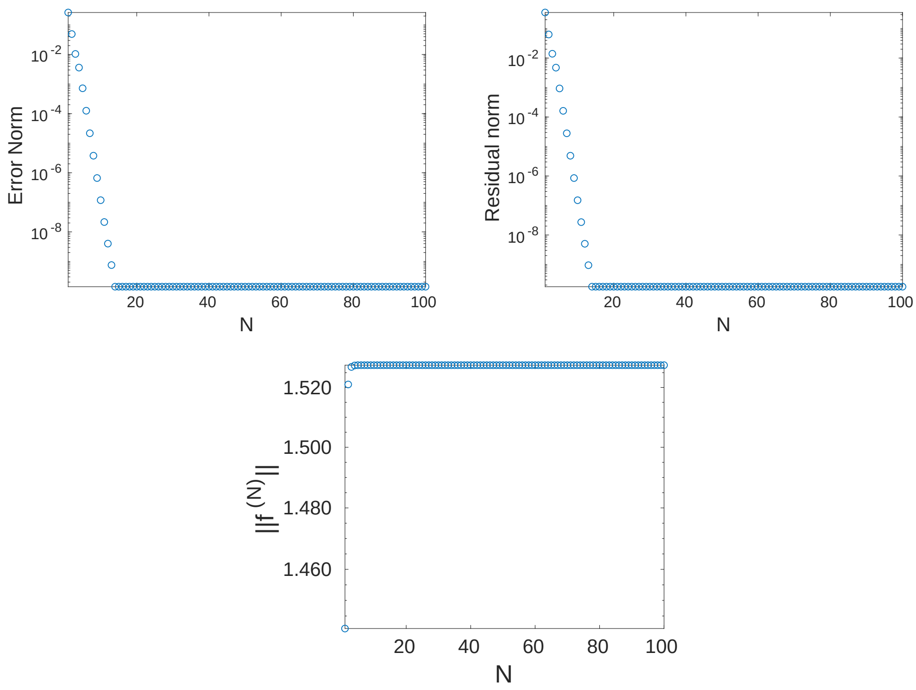
(c) Krylov basis truncation

Figure 3.2: Norm of the infinite-dimensional error and residual, and of the approximated solution, for the Volterra inverse problem truncated with the Legendre, complex Fourier, and Krylov bases.



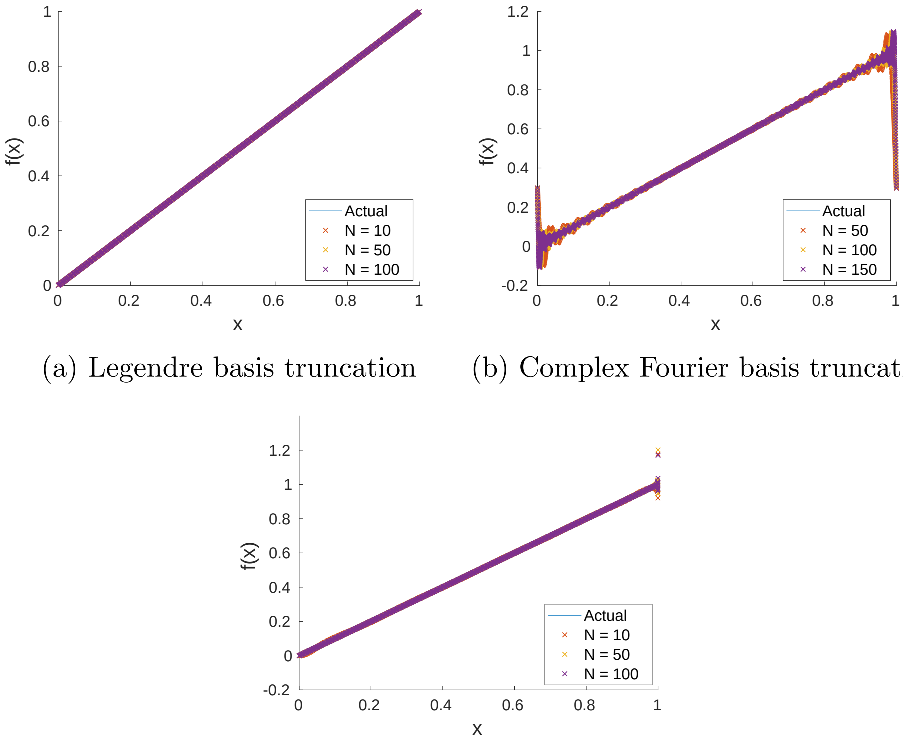
(a) Legendre basis truncation

(b) Complex Fourier basis truncation



(c) Krylov basis truncation

Figure 3.3: Norm of the infinite-dimensional error and residual, and of the approximated solution, for the M-multiplication inverse problem truncated with the Legendre, complex Fourier, and Krylov bases.

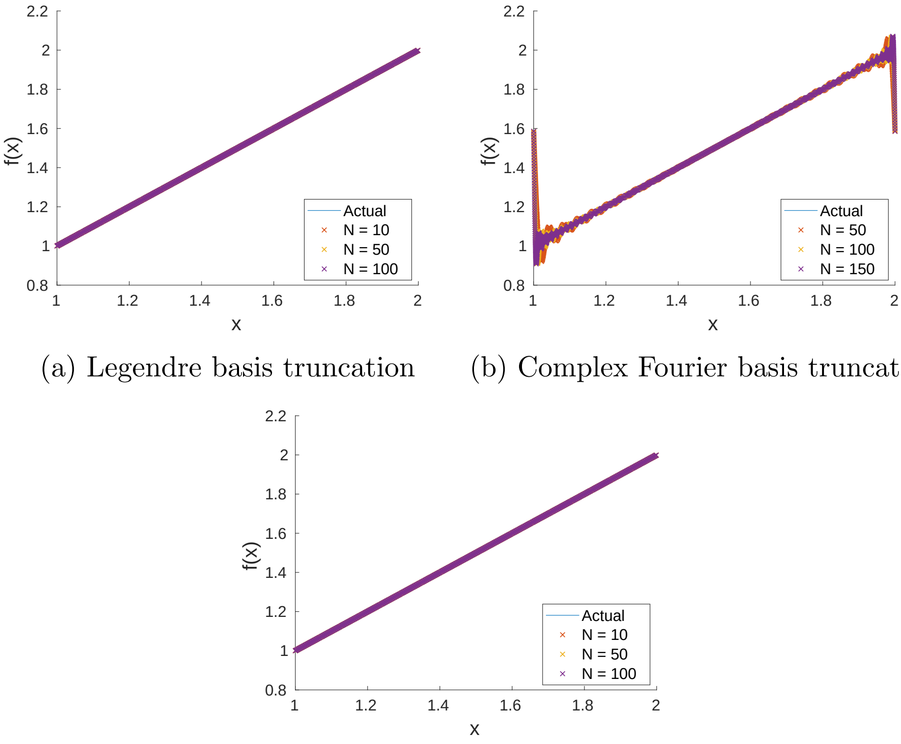


(a) Legendre basis truncation

(b) Complex Fourier basis truncation

(c) Krylov basis truncation

Figure 3.4: Reconstruction of the exact solution $f_1(x) = x$ from the approximate solutions for the problem $Vf_1 = g_1$. The complex Fourier basis produces an incaccurate reconstruction due to high oscillations, resulting in higher errors.



(a) Legendre basis truncation

(b) Complex Fourier basis truncation

(c) Krylov basis truncation

Figure 3.5: Reconstruction of the exact solution $f_2(x) = x$ from the approximate solutions for the problem $Mf_2 = g_2$. Again, the complex Fourier basis produces the least accurate reconstruction.

Chapter 4

Krylov Solutions in Hilbert Space for Bounded Inverse Problems

4.1 Introduction

Krylov subspace methods are some of the most popular algorithms in numerical analysis, especially due to their speed. The framework surrounding Krylov methods in finite-dimensions is a well-studied and deeply understood area. Although there is some analysis for these methods in infinite-dimensions (see Chapter 2 for an overview), currently there is no *systematic* study. In fact, many of the studies in infinite-dimensions concern particular classes of operator equations and particular methods. A nice example is the conjugate-gradient method. Although this method is restricted to the class of self-adjoint, bounded, positive operators on Hilbert space; it is known *always* to converge strongly to a solution to the linear inverse problem [64].

This Chapter, based on the work [16], focuses on the infinite-dimensional setting for the solution to inverse linear problems using Krylov subspace methods. Operator theoretic notions, with necessary and sufficient conditions, are developed to ensure that a solution to the linear inverse problem is arbitrarily well approximated by vectors in the Krylov subspace.

The problem to be considered in this Chapter is the *linear inverse problem* on a Hilbert space $\mathcal{H}$ where $\dim \mathcal{H} = \infty$. Recall that

$$Af = g\,, \tag{4.1}$$

for $A \in \mathscr{B}(\mathcal{H})$, $g \in \mathcal{H}$ a vector, and $f \in \mathcal{H}$ a (possible) solution(s) to the problem. At this point it is stressed that A is a *bounded* linear operator. Unbounded operators are considered later in Chapters 5 and 6. Recall that (4.1) is called: *solvable* if there exists some $f \in \mathcal{H}$ that satisfies (4.1) (i.e. $g \in \mathrm{ran}A$); *well-defined* if additionally the solution is unique (i.e. A is injective); and *well-posed* if there exists a unique $f \in \mathcal{H}$ satisfying (4.1) that depends continuously on the datum g (i.e. A has an everywhere defined bounded inverse).

Krylov subspace methods use linear combinations of the vectors $g, Ag, A^2g, \ldots$ spanning the Krylov space $\mathcal{K}(A,\, g)$ to approximate solutions to the linear inverse problem (4.1). If a solution $f \in \mathcal{H}$ to (4.1) can be arbitrarily well approximated by linear combinations of these vectors, then f is called a *Krylov solution* and the problem (4.1) is termed *Krylov-solvable*. That is to say, $f \in \mathcal{H}$ is a Krylov solution to the solvable (4.1) if $f \in \overline{\mathcal{K}(A,\, g)}$, where the closure is taken in the $\mathcal{H}$-norm topology on $\mathcal{H}$.

Although aspects of Krylov solutions and Krylov-solvability are trivial in the finite-dimensional setting, this theory is not so obvious once one moves to general Hilbert spaces. For example, $\mathcal{K}(A,\, g)$ may not be dense in the ambient Hilbert space $\mathcal{H}$, and as such the approximability characteristic in the Petrov-Galerkin projection method may be lost. Moreover, the question as to the uniqueness of solution(s), should they exist, in the Krylov space is important.

This Chapter begins with some formal definitions of the Krylov space and general comments. Then the important operator-theoretic aspects of Krylov reducibility and Krylov intersection are introduced, along with some examples of Krylov solvability (or lack thereof), and finally general conditions for Krylov solvability are considered. The theory developed here is supported with some simple numerical tests.

4.2 Definitions and comments

In *finite-dimensional* space $\mathbb{C}^m$ with a matrix $A \in \mathbb{C}^{m\times m}$ and vector $g \in \mathbb{C}^m$, the N-th order Krylov space associated with A and g is given by

$$\mathcal{K}_N\left(A,\, g\right) := \operatorname{span}\left\{g, Ag, A^2 g, \ldots, A^{N-1}g\right\} .$$

Clearly, $1 \leq \dim \mathcal{K}_N\left(A,\, g\right) \leq m$, and there always exists some $m_0 \leq m$ such that $\mathcal{K}_{m_0}\left(A,\, g\right) = \mathcal{K}_N\left(A,\, g\right)$ for all $N \geq m_0$. This idea is now extended to the infinite-dimensional setting, where the relevant terminology is still applicable in the finite-dimensional setting.

Definition 4.2.1. Let $\mathcal{H}$ be a Hilbert space, let $A : \mathcal{H} \to \mathcal{H}$ be a bounded linear operator on $\mathcal{H}$, and consider some $g \in \mathcal{H}$. Then the N-th order Krylov subspace associated with A and g is

$$\mathcal{K}_N\left(A,\, g\right) := \operatorname{span}\left\{A^n g \,|\, n \in \mathbb{N}_0,\, n \leq N-1\right\} , \tag{4.2}$$

and *the* Krylov subspace associated with A and g is defined as

$$\mathcal{K}\left(A,\, g\right) := \operatorname{span}\left\{A^n g \,|\, n \in \mathbb{N}_0\right\} , \tag{4.3}$$

where $\operatorname{span}\{\cdot\}$ refers to the set of *finite* linear combinations of its arguments.

One always has that $\sup_N \dim \mathcal{K}_N\left(A,\, g\right) = \dim \mathcal{K}\left(A,\, g\right)$, however $\dim \mathcal{K}\left(A,\, g\right) = \infty$ is possible when $\dim \mathcal{H} = \infty$. When one has that $\dim \mathcal{K}\left(A,\, g\right) = \infty$, it is evident that the Krylov space is not closed, but also not open. The closure of $\mathcal{K}\left(A,\, g\right)$ may be a proper subspace of $\mathcal{H}$ or the entire Hilbert space.

Example 4.2.2. (i) The right shift operator $\mathcal{R} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ (see Appendix A) with the vector $g = e_{m+1}$ for some $m \in \mathbb{Z}$, generates the Krylov space

$$\overline{\mathcal{K}\left(\mathcal{R},\, g\right)} = \overline{\operatorname{span}\left\{e_{m+1}, e_{m+2}, \ldots\right\}} = \operatorname{span}\left\{e_n;\, n \leq m\right\}^{\perp}$$

which is *always* a proper subspace of $\ell^2(\mathbb{Z})$.

(ii) The right shift operator $R : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ with the vector $g = e_{m+1}$ for some $m \in \mathbb{N}$, generates the Krylov spaces as mentioned above. Here, $\overline{\mathcal{K}(A,\, g)} = \mathcal{H}$ when one has that $g = e_1$.

(iii) Consider the Volterra operator $V : L^2[0,1] \to L^2[0,1]$, $f(x) \mapsto \int_0^x f(y)\,\mathrm{d}y$ (see Appendix A), and the constant function $g = \mathbf{1}$. Then it follows that the powers of V applied to g give multiples of the polynomials, so that

$$\mathcal{K}(V,\, \mathbf{1}) = \operatorname{span}\left\{1, x, x^2, \ldots\right\} .$$

is the space of all polynomials on $[0,1]$. A consequence of the Stone-Weierstrass theorem (Theorem C.2.3) shows that one has that the space of polynomials on $[0,1]$ is dense in $L^2[0,1]$, so therefore $\overline{\mathcal{K}(V,\, \mathbf{1})} = L^2[0,1]$.

These examples highlight an interesting theoretical concept, namely that of *cyclicity.* In purely operator-theoretical terms, the Krylov subspace $\mathcal{K}(A,\, g)$ is referred to as the *cyclic space* of A relative to the vector g, and $g, Ag, A^2g, \ldots$ form the *orbit* of g under A. Density of $\mathcal{K}(A,\, g)$ in the ambient Hilbert space is called the *cyclicity* of the vector g, in which case one calls g a *cyclic vector* and A a *cyclic operator.*

Remark 4.2.3. Interesting properties and well known results of cyclic operators are summarised in the monograph by Halmos [40], and are listed here for completeness.

1. In non-separable Hilbert space, there are no cyclic vectors.

2. If the operator $A \in \mathscr{B}(\mathcal{H})$ is non-scalar and commutes with a backward (i.e., left) shift on $\mathcal{H}$, then A is cyclic with an A-invariant, dense vector manifold of cyclic vectors [36].

3. The set of bounded cyclic operators is *not* dense in $\mathscr{B}(\mathcal{H})$ with respect to $\|\cdot\|_{\text{op}}$ when $\dim \mathcal{H} = \infty$. On the other hand, when $\dim \mathcal{H} < \infty$, then the set of bounded cyclic operators *is* dense in $\mathscr{B}(\mathcal{H})$ with respect to $\|\cdot\|_{\text{op}}$.

4. The set of cyclic operators is open in $\mathscr{B}(\mathcal{H})$ when $\dim \mathcal{H} < \infty$; and when $\dim \mathcal{H} = \infty$, the set of cyclic operators is not closed.

5. If $\dim \mathcal{H} = \infty$ and $\mathcal{H}$ is separable, then the set of non-cyclic operators on $\mathcal{H}$ is dense in $\mathscr{B}(\mathcal{H})$.

6. It is unknown whether there exists a cyclic operator on a separable Hilbert space $\mathcal{H}$ such that *every* non-trivial vector in $\mathcal{H}$ is a cyclic vector. Such an operator would be a counter example to the famous invariant subspace problem (see Remark 4.3.2).

7. The set of cyclic vectors for a bounded linear operator A on $\mathcal{H}$ is either empty or a *dense* subset of $\mathcal{H}$ [34]. If $v \in \mathcal{H}$ is a cyclic vector of A, then so too is $v^{(n)} = (\mathbb{1} - \alpha A)^n v$, for all $|\alpha| \in (0, \|A^{-1}\|_{\mathrm{op}})$ and for all $n \in \mathbb{N}$. Clearly then, the $v^{(n)}$'s span $\mathcal{H}$.

8. A bounded linear operator $A \in \mathscr{B}(\mathcal{H})$ is cyclic if and only if there exists some orthonormal basis $(e_n)_{n\in\mathbb{N}}$ such that $\langle e_i, Ae_j\rangle$ is non-zero for $j = i + 1$, and zero for all $i > j + 1$.

4.3 Krylov reducibility and Krylov intersection

The concepts of Krylov reducibility and Krylov intersection are some of the *fundamental* operator-theoretic mechanisms that are studied in this, and the next, chapter. For a given operator $A \in \mathscr{B}(\mathcal{H})$ and $g \in \mathcal{H}$, there exists the orthogonal decomposition [10, Chapter 5]

$$\mathcal{H} = \overline{\mathcal{K}(A,\, g)} \oplus \mathcal{K}(A,\, g)^{\perp} \ , \tag{4.4}$$

that is referred to as the *Krylov decomposition* of $\mathcal{H}$ relative to A and g. It is immediate from the definition of the Krylov subspace that it is invariant under the action of A. In addition, the space $\mathcal{K}(A,\, g)^{\perp}$ is closed and invariant under the action of the adjoint operator A^*. One has the following lemma.

Lemma 4.3.1. *Given $A \in \mathscr{B}(\mathcal{H})$ and $g \in \mathcal{H}$, the following results regarding the invariance of the Krylov subspace and its perpendicular complement hold*

$$A\overline{\mathcal{K}(A,\,g)} \subset \overline{\mathcal{K}(A,\,g)}\,, \quad A^*\mathcal{K}(A,\,g)^{\perp} \subset \mathcal{K}(A,\,g)^{\perp}\,. \tag{4.5}$$

Proof. The inclusion $A\mathcal{K}(A,\,g) \subset \mathcal{K}(A,\,g)$ is obvious from the definition of the Krylov subspace.

For any continuous function between topological spaces X and Y, $h : X \to Y$ and any subset $\mathcal{V} \subset X$, one has [59, Chapter 2, Section 18]

$$h(\mathcal{V}) \subset h(\overline{\mathcal{V}}) \subset \overline{h(\mathcal{V})}\,, \tag{4.6}$$

and moreover if h is a homeomorphism

$$h(\overline{\mathcal{V}}) = \overline{h(\mathcal{V})}\,. \tag{4.7}$$

The first inclusion in (4.5) immediately follows from (4.6). When $h = A$ and $X = Y = \mathcal{H}$, the conditions for (4.6) to be true merely require $A \in \mathscr{B}(\mathcal{H})$. While for (4.7) to be true, one additionally requires that $A^{-1} \in \mathscr{B}(\mathcal{H})$.

The second inclusion follows from the fact that $\langle A^*w,\, z\rangle = \langle w,\, Az\rangle = 0$ for all $z \in \mathcal{K}(A,\,g)$, where w is any vector in $\mathcal{K}(A,\,g)^{\perp}$. Taking limiting sequences of vectors in $\mathcal{K}(A,\,g)$, one has $\langle A^*w,\, z\rangle = \langle w,\, Az\rangle = 0$ for all $z \in \overline{\mathcal{K}(A,\,g)}$, where $w \in \mathcal{K}(A,\,g)^{\perp}$. □

Remark 4.3.2. The invariance of the Krylov subspace under the action of the operator A is in some sense related to the famous invariant subspace problem [6]. The question asks whether every $A \in \mathscr{B}(\mathcal{H})$ has a non-trivial, i.e., neither $\{0\}$ nor $\mathcal{H}$, closed invariant subspace $\mathcal{V} \subset \mathcal{H}$. On Hilbert spaces the answer to this question is still unknown.

The next concept is one that is core in discussing the Krylov-solvability of certain classes of problems, namely that of *Krylov reducibility.*

Definition 4.3.3. Given $A \in \mathscr{B}(\mathcal{H})$ and $g \in \mathcal{H}$, one says that the operator A is *reduced* by the Krylov decomposition (4.4), or A is $\mathcal{K}(A,\,g)$-reduced, if

$\overline{\mathcal{K}(A,\, g)}$ *and* $\mathcal{K}(A,\, g)^{\perp}$ are invariant under A. This feature of A is referred to as $\mathcal{K}(A,\, g)$*-reducibility*, or *Krylov reducibility* where no confusion arises.

It is immediate that if A is $\mathcal{K}(A,\, g)$-reduced, then so is A^*, and visa-versa.

Lemma 4.3.4. *If* $A \in \mathscr{B}(\mathcal{H})$ *and* $\mathcal{V} \subset \mathcal{H}$ *is a* closed *subspace of* $\mathcal{H}$*, then (i) and (ii) below are equivalent:*

(i) $A\mathcal{V} \subset \mathcal{V}$ *and* $A\mathcal{V}^{\perp} \subset \mathcal{V}^{\perp}$*,*

(ii) $A^*\mathcal{V} \subset \mathcal{V}$ *and* $A^*\mathcal{V}^{\perp} \subset \mathcal{V}^{\perp}$*.*

Proof. Assume that (i) is true. Then take any $w \in \mathcal{V}^{\perp}$ and any $z \in \mathcal{V}$. Immediately, $0 = \langle w,\, Az\rangle = \langle A^*w,\, z\rangle$. So $A^*w \perp z$ for all $z \in \mathcal{V}$. As $w \in \mathcal{V}^{\perp}$ is arbitrary, it follows that $A^*\mathcal{V}^{\perp} \subset \mathcal{V}^{\perp}$. Similarly, one has $0 = \langle z,\, Aw\rangle = \langle A^*z,\, w\rangle$ and by the same argument as above, $A^*\mathcal{V} \subset \mathcal{V}$.

The converse statement that (ii) $\implies$ (i) is similar, and can be seen by interchanging the role of A and A^*. □

Remark 4.3.5. For a general $A \in \mathscr{B}(\mathcal{H})$ and $g \in \mathcal{H}$, A may not be $\mathcal{K}(A,\, g)$-reduced (see Example 4.3.9), however *all bounded self-adjoint operators* A *are* $\mathcal{K}(A,\, g)$*-reduced* due to (4.5) and Lemma 4.3.4. This Krylov reducibility feature is not restricted just to the class of self-adjoint operators as the next example reveals.

Example 4.3.6. Consider two operators $A, B \in \mathscr{B}(\mathcal{H})$, and define a new operator $\widetilde{A} : \widetilde{\mathcal{H}} \to \widetilde{\mathcal{H}}$ on the Hilbert space $\widetilde{\mathcal{H}} = \mathcal{H} \oplus \mathcal{H}$, with $\widetilde{A} := A \oplus B$. Let $g \in \mathcal{H}$ be a cyclic vector for A in $\mathcal{H}$, and take $\widetilde{g} := g \oplus 0$. Then $\overline{\mathcal{K}\left(\widetilde{A},\, \widetilde{g}\right)} = \mathcal{H} \oplus \{0\}$, so $\mathcal{K}\left(\widetilde{A},\, \widetilde{g}\right)^{\perp} = \{0\} \oplus \mathcal{H}$. Therefore, $\widetilde{A}$ is $\mathcal{K}\left(\widetilde{A},\, g\right)$-reduced, and yet $\widetilde{A}$ is self-adjoint on $\widetilde{\mathcal{H}}$ if and only if *both* A and B are self-adjoint on $\mathcal{H}$.

The characterisation of Krylov reducibility may be a non-trivial task for general operators. *Normal operators* on the other hand, have the following equivalent characterisation of the Krylov reducibility.

Proposition 4.3.7. *Let* $A \in \mathscr{B}(\mathcal{H})$ *be a normal operator, and* $g \in \mathcal{H}$*. Then* A *is* $\mathcal{K}(A,\, g)$*-reduced if and only if* $A^*g \in \overline{\mathcal{K}(A,\, g)}$*.*

Proof. If A is $\mathcal{K}(A,\, g)$-reduced, then $\overline{\mathcal{K}(A,\, g)}$ is invariant under A^* (Lemma 4.3.4). In particular, $A^*g \in \overline{\mathcal{K}(A,\, g)}$.

Conversely, let $A^*g \in \overline{\mathcal{K}(A,\, g)}$. Then from Lemma 4.3.1,

$$\overline{\mathcal{K}(A,\, A^*g)} = \overline{\operatorname{span}\{A^nA^*g \,|\, n \in \mathbb{N}_0\}} \subset \overline{\mathcal{K}(A,\, g)}\,,$$

and since A is normal, $A^*\mathcal{K}(A,\, g) = \mathcal{K}(A,\, A^*g)$; so using (4.6)

$$A^*\overline{\mathcal{K}(A,\, g)} \subset \overline{\mathcal{K}(A,\, A^*g)} \subset \overline{\mathcal{K}(A,\, g)}\,.$$

This property together with (4.5) implies that A^* is $\mathcal{K}(A,\, g)$-reduced, and therefore so is A (Lemma 4.3.4). □

Following these results, a core concept known as the *Krylov intersection* is defined below. As shall be seen, this is the operator-theoretic notion that captures the *essence* of Krylov-solvability in a general sense.

Definition 4.3.8. Given a bounded linear operator A on Hilbert space $\mathcal{H}$ and a vector $g \in \mathcal{H}$, the intersection

$$\overline{\mathcal{K}(A,\, g)} \cap (A\mathcal{K}(A,\, g)^{\perp})\,, \tag{4.8}$$

is called the *Krylov intersection* with respect to A and g, and is denoted by $\mathscr{I}_{\mathcal{K}}(A,\, g)$.

For $A \in \mathscr{B}(\mathcal{H})$ and $g \in \mathcal{H}$, a consequence of A being Krylov reducible is that the Krylov intersection is trivial, i.e., $\mathscr{I}_{\mathcal{K}}(A,\, g) = \{0\}$. The converse is not true in general.

Example 4.3.9. The Krylov intersection may still be trivial, even in the absence of Krylov reducibility. This information is immediately clear even just at the finite-dimensional level for matrices. For example, taking in the Hilbert space $\mathbb{C}^2$

$$A_\theta = \begin{pmatrix} 1 & \cos\theta \\ 0 & \sin\theta \end{pmatrix} \qquad \theta \in (0, \tfrac{\pi}{2}]\,, \qquad g = \begin{pmatrix} 1 \\ 0 \end{pmatrix}\,,$$

one sees that A_θ is $\mathcal{K}(A_\theta,\, g)$-reduced only in the case when $\theta = \frac{\pi}{2}$, while $\mathscr{I}_{\mathcal{K}}(A_\theta,\, g) = \{0\}$ for any $\theta \in (0, \frac{\pi}{2}]$.

4.4 Krylov Solvability

In this Section we revisit the linear inverse problem (4.1) and consider the question of Krylov solvability. Given some $A \in \mathscr{B}(\mathcal{H})$ and $g \in \operatorname{ran}A$, one searches for solution(s) $f \in \mathcal{H}$ to (4.1); more specifically one asks: when does a solution f to $Af = g$ admit arbitrarily close approximants in the space $\mathcal{K}(A,\, g)$? These approximants are, of course, formed by finite linear combinations of vectors in $\mathcal{K}(A,\, g)$, and so are suitable for numerical calculations. Therefore one would like that $f \in \overline{\mathcal{K}(A,\, g)}$ to have a sound numerical scheme.

Recall that a solution $f \in \mathcal{H}$ to (4.1) belonging to the space $\overline{\mathcal{K}(A,\, g)}$ is called a *Krylov solution*, and a linear inverse problem that has solution(s) with this property is called *Krylov solvable*. Informally, one may use the expression *Krylov solvability* to describe a linear inverse problem that exhibits such solution(s).

4.4.1 Examples of Krylov solvability (or lack of)

Although the examples presented here require an analysis specific to the chosen operator A and vector g in (4.1), they serve an informative purpose in unmasking the general theory.

Example 4.4.1. (i) The self-adjoint multiplication operator $M_x : L^2[1,2] \to L^2[1,2]$ with action $\phi \mapsto x\phi$ is bounded and invertible with an everywhere defined bounded inverse $M_x^{-1} : L^2[1,2] \to L^2[1,2]$, $\phi \mapsto \frac{1}{x}\phi$. The solution to $M_x f = \mathbf{1}$ is the function $f(x) = \frac{1}{x}$. One has $\mathcal{K}(M_x,\, \mathbf{1}) = \{p(x) \,|\, p \in \mathbb{P}_{[1,2]}[x]\}$, where $\mathbb{P}_{[1,2]}[x]$ denotes the space of polynomials on the domain $[1,2]$. As the polynomials are *dense* on $L^2[1,2]$ (a consequence of Theorem C.2.3), immediately one has that f is a Krylov solution.

(ii) The multiplication operator $M_z : L^2(\Omega) \to L^2(\Omega)$, $f \mapsto zf$ on the domain $\Omega := \{z \in \mathbb{C};\, |z - \frac{3}{4}| < \frac{1}{4}\}$ (Appendix A) is a normal, bounded bijection on $L^2(\Omega)$. The solution f to $M_z f = g$ for a given $g \in L^2(\Omega)$ is the function $f(z) = z^{-1}g(z)$, so M_z^{-1} is the map defined by $g \mapsto z^{-1}g$ for $g \in L^2(\Omega)$. The Krylov space is given by $\mathcal{K}(M_z, g) = \{pg \,|\, p \in \mathbb{P}_\Omega[z]\}$ where $\mathbb{P}_\Omega[z]$ denotes a polynomial in z on Ω. Certainly one has that z^{-1} is holomorphic in Ω and that there exists a power series (e.g., the Taylor expansion of z^{-1} about $z = \frac{3}{4}$) that converges uniformly to z^{-1} on Ω [79]. In fact, $f \in \overline{\mathcal{K}(M_z, g)}$ and hence the problem $M_z f = g$ is *always* Krylov solvable. Indeed, choosing a sequence of L^∞-approximants among the polynomials on Ω, $(p_n)_n$, one has

$$\begin{aligned}\|f - p_n g\|_{L^2(\Omega)} &= \left\|(z^{-1} - p_n)g\right\|_{L^2(\Omega)} \\ &\leq \left\|z^{-1} - p_n\right\|_\infty \|g\|_{L^2(\Omega)} \xrightarrow{n\to\infty} 0\,.\end{aligned}$$

(iii) The left-shift operator L on $\ell^2(\mathbb{N}_0)$ (Appendix A) is bounded, non-injective, with range $\mathrm{ran}L = \ell^2(\mathbb{N}_0)$. In fact, this operator is cyclic owing to the properties described in Remark 4.2.3. The solution to $Lf = g$ with $g := \sum_{n\in\mathbb{N}_0} \frac{1}{n!}e_n$ is $f = \sum_{n\in\mathbb{N}_0} \frac{1}{n!}e_{n+1}$. $\mathcal{K}(L, g)$ is *dense* in $\ell^2(\mathbb{N}_0)$ so clearly f is a Krylov solution. To reveal the density of $\mathcal{K}(L, g)$, one may see that the vector $e_0 \in \overline{\mathcal{K}(L, g)}$ because

$$\begin{aligned}&\left\|k!L^k g - e_0\right\|^2_{\ell^2(\mathbb{N}_0)} \\ &\qquad = \left\|(1, \frac{1}{k+1}, \frac{1}{(k+2)(k+1)}, \cdots) - (1, 0, 0, \cdots)\right\|^2_{\ell^2(\mathbb{N}_0)} \\ &\qquad = \sum_{n=1}^{\infty} \left(\frac{k!}{(n+k)!}\right)^2 \xrightarrow{k\to\infty} 0\,.\end{aligned}$$

So, $(0, \frac{1}{k!}, \frac{1}{(k+1)!}, \cdots) = L^{k-1}g - (k-1)!e_0 \in \overline{\mathcal{K}(L, g)}$, and so the vector e_1 also belongs to $\overline{\mathcal{K}(L, g)}$ because

$$\left\|k!(L^{k-1}g - (k-1)!e_0) - e_1\right\|^2_{\ell^2(\mathbb{N}_0)} = \sum_{n=1}^{\infty} \left(\frac{k!}{(n+k)!}\right)^2 \xrightarrow{k\to\infty} 0\,.$$

This argument can then be repeated for any e_n by induction, so that $e_n \in \overline{\mathcal{K}(L,\, g)}$ for all $n \in \mathbb{N}_0$.

(iv) The right-shift operator R on $\ell^2(\mathbb{N})$ (see Appendix A) is bounded and injective, with non-dense range. The solution to $Rf = e_2$ is $f = e_1$. But, f is *not* a Krylov solution, as $\overline{\mathcal{K}(R,\, e_2)} = \overline{\operatorname{span}\{e_2, e_3, \ldots\}}$. Therefore the problem $Rf = e_2$ is not Krylov solvable.

(v) The compact, or weighted, right-shift operator $\mathcal{R}$ on $\ell^2(\mathbb{Z})$ (see Appendix A) is normal, injective, and has dense range. The solution to $\mathcal{R}f = \sigma_1 e_2$ is $f = e_1$. However, f is *not* a Krylov solution, as $\overline{\mathcal{K}(\mathcal{R},\, e_2)} = \overline{\operatorname{span}\{e_2, e_3, \ldots\}}$. Again, the problem $\mathcal{R}f = \sigma_1 e_1$ is not Krylov solvable. The same may also be said about the unweighted right-shift of Example 4.2.2.

(vi) Let A be a bounded injective operator on a Hilbert space $\mathcal{H}$ with cyclic vector $g \in \operatorname{ran}A$. Let $\varphi_0 \in \mathcal{H} \setminus \{0\}$, and let $f \in \mathcal{H}$ be the solution to $Af = g$. Consider the Hilbert space $\widetilde{\mathcal{H}} = \mathcal{H} \oplus \mathcal{H}$ and the operator $\widetilde{A} := A \oplus |\varphi_0\rangle\langle\varphi_0|$. Obviously $\widetilde{A}$ has an infinite-dimensional kernel $\ker \widetilde{A} = \{0\} \oplus \operatorname{span}\{\varphi_0\}^{\perp}$. One possible solution to the problem $\widetilde{A}\widetilde{f} = \widetilde{g} := g \oplus 0$ is $\widetilde{f} = f \oplus 0$, and $\widetilde{f} \in \mathcal{H} \oplus \{0\} = \overline{\mathcal{K}\left(\widetilde{A},\, \widetilde{g}\right)}$. Another possibility is that $\widetilde{f}_\xi = f \oplus \xi$ where $\xi \in \mathcal{H} \setminus \{0\}$ and $\xi \perp \varphi_0$. Obviously, $\widetilde{f}_\xi \notin \overline{\mathcal{K}\left(\widetilde{A},\, \widetilde{g}\right)}$. This operator therefore exhibits Krylov solutions, but also an infinite amount of solutions that are *not* in the closed Krylov space.

(vii) If V is the Volterra operator on $L^2[0,1]$ (Appendix A) and $g(x) = \frac{1}{2}x^2$, then $f(x) = x$ is the unique solution to $Vf = g$. Considering the Krylov space $\mathcal{K}(V,\, g)$, it is spanned by the monomials $x^2, x^3, \ldots$, from which

$$\mathcal{K}(V,\, g) = \{x^2 p(x) \,|\, p \in \mathbb{P}_{[0,1]}[x]\}\,.$$

Clearly then, $f \notin \mathcal{K}(V,\, g)$, as $f(x) = x^2 \cdot \frac{1}{x}$ and $\frac{1}{x} \notin L^2[0,1]$. But interestingly, $\mathcal{K}(V,\, g)$ is *dense* in $L^2[0,1]$, so that $f \in \overline{\mathcal{K}(V,\, g)}$. Indeed, consider some $h \in \mathcal{K}(V,\, g)^{\perp}$, then $\int_0^1 \overline{h(x)}x^2 p(x)\,\mathrm{d}x = 0$ for any polynomial p. The L^2 density of polynomials on $[0,1]$ implies that $x^2 h(x) = 0$ a.e., from which $h = 0$ a.e.. This shows that $\mathcal{K}(V,\, g)^{\perp} = \{0\}$ and hence $\overline{\mathcal{K}(V,\, g)} = L^2[0,1]$.

4.4.2 General conditions for Krylov solvability

Example 4.4.1 reveals that even stringent assumptions on the operator A, such as the simultaneous occurrence of normality, injectivity, density of the range, compactness, or even bounded invertibility, do *not* guarantee, in general, that the solution to $Af = g$, for $g \in \mathrm{ran}A$, is a Krylov solution. This is quite contrary to the finite-dimensional situation, whereby the invertibility alone of the (matrix) operator $A \in \mathbb{C}^{m\times m}$ is enough to guarantee $f \in \mathcal{K}_m(A,\, g)$ for the linear inverse problem.

A *necassary* condition for the solution to a well-defined linear inverse problem, that becomes necessary and sufficient if A is a bounded bijection (i.e. a homeomorphism), is stated in the following proposition.

Proposition 4.4.2. *Let A be a bounded and injective operator on a Hilbert space $\mathcal{H}$, and let f be the solution to $Af = g$, given $g \in \mathrm{ran}A$. One has the following.*

(i) If $f \in \overline{\mathcal{K}(A,\, g)}$, then $A\overline{\mathcal{K}(A,\, g)}$ is dense in $\overline{\mathcal{K}(A,\, g)}$.

(ii) Assume further that A is invertible with an everywhere defined, bounded inverse on $\mathcal{H}$. Then $f \in \overline{\mathcal{K}(A,\, g)}$ if and only if $A\overline{\mathcal{K}(A,\, g)}$ is dense in $\overline{\mathcal{K}(A,\, g)}$.

Proof. It is obvious that $A\overline{\mathcal{K}(A,\, g)} \supset A\mathcal{K}(A,\, g) = \mathrm{span}\left\{A^k g \,|\, k \in \mathbb{N}_0\right\}$, owing to definition 4.2.1 and (4.6). If $f \in \overline{\mathcal{K}(A,\, g)}$, then $Af = g \in A\overline{\mathcal{K}(A,\, g)}$ so that one has $A\overline{\mathcal{K}(A,\, g)} \supset \mathrm{span}\left\{A^k g \,|\, k \in \mathbb{N}_0\right\}$; the latter implying that $\overline{\mathcal{K}(A,\, g)} \supset \overline{A\overline{\mathcal{K}(A,\, g)}} \supset \overline{\mathcal{K}(A,\, g)}$ by (4.6) and (4.5), from which one has $\overline{A\overline{\mathcal{K}(A,\, g)}} = \overline{\mathcal{K}(A,\, g)}$. This proves part (i), and the 'only if' implication in part (ii).

For the converse in part (ii), consider that $A^{-1} \in \mathscr{B}(\mathcal{H})$ and that $A\overline{\mathcal{K}(A,\, g)}$ is dense in $\overline{\mathcal{K}(A,\, g)}$. Let $(Av_n)_{n\in\mathbb{N}}$ be a sequence in $A\overline{\mathcal{K}(A,\, g)}$ that tends to

$g \in \overline{\mathcal{K}(A,\, g)}$, for some v_n's in $\overline{\mathcal{K}(A,\, g)}$. Since A^{-1} is bounded on $\mathcal{H}$, one has that $(v_n)_{n\in\mathbb{N}}$ is a Cauchy sequence, as $(Av_n)_{n\in\mathbb{N}}$ is Cauchy so $\|v_n - v_m\|_{\mathcal{H}} \leq \|A^{-1}\|_{\mathrm{op}} \|Av_n - Av_m\|_{\mathcal{H}} \to 0$ as $n, m \to \infty$. Therefore, $v_n \to v \in \overline{\mathcal{K}(A,\, g)}$ as $n \to \infty$. From continuity, $Af = g = \lim_{n\to\infty} Av_n = Av$, and by injectivity one has that $f = v \in \overline{\mathcal{K}(A,\, g)}$. □

A *sufficient* condition to ensure Krylov solvability of the well-defined linear inverse problem is that A is $\mathcal{K}(A,\, g)$-reduced.

Proposition 4.4.3. *Let A be a bounded and injective operator on a Hilbert space $\mathcal{H}$, and let $f \in \mathcal{H}$ be the solution to $Af = g$, given $g \in \mathrm{ran}A$. If A is $\mathcal{K}(A,\, g)$-reduced, then $f \in \overline{\mathcal{K}(A,\, g)}$. In particular, if A is bounded, injective and self-adjoint, then $Af = g$ implies $f \in \overline{\mathcal{K}(A,\, g)}$.*

Proof. Let $P_{\mathcal{K}} : \mathcal{H} \to \mathcal{H}$ be the orthogonal projection onto $\overline{\mathcal{K}(A,\, g)}$. Immediately $Af = g \in \overline{\mathcal{K}(A,\, g)}$ and owing to the invariance relation (4.5) one has $AP_{\mathcal{K}}f \in \overline{\mathcal{K}(A,\, g)}$. As $Af = g = AP_{\mathcal{K}}f + A(\mathbb{1} - P_{\mathcal{K}})f$, it immediately follows that $A(\mathbb{1} - P_{\mathcal{K}})f \in \overline{\mathcal{K}(A,\, g)}$.

But, owing to A being $\mathcal{K}(A,\, g)$-reduced, one has $A(\mathbb{1} - P_{\mathcal{K}})f \in \mathcal{K}(A,\, g)^{\perp}$. Necessarily, $A(\mathbb{1} - P_{\mathcal{K}})f = 0$ and by injectivity $f = P_{\mathcal{K}}f \in \overline{\mathcal{K}(A,\, g)}$. By Remark 4.3.5 the self-adjoint case immediately follows. □

Krylov solvability for bounded, self-adjoint operators can be concluded through the following alternative route.

Proposition 4.4.4. *Let A be a bounded, self-adjoint operator on a Hilbert space $\mathcal{H}$ with spectrum $\sigma(A)$. Let $\mathbf{E}(t)$ be the spectral measure for A, with the associated scalar measure $\mu_g(t) := \langle g,\, \mathbf{E}(t)\, g\rangle$ associated to a given $g \in \mathrm{ran}A$. Then for any $h \in L^2(\sigma(A), \mu_g)$ one has that $h(A)g \in \overline{\mathcal{K}(A,\, g)}$. In addition, if A is injective, then the solution f to $Af = g$ for $g \in \mathrm{ran}A$ is in $\overline{\mathcal{K}(A,\, g)}$.*

Proof. Observe preliminarily that $\sigma(A) \subset [-\|A\|_{\mathrm{op}}, \|A\|_{\mathrm{op}}]$ and that μ_g is positive (as $\langle g,\, \mathbf{E}(t)\, g\rangle \geq 0$) and regular (as $\mu_g(K) < \infty$ for every compact $K \subset \mathbb{R}$).

By standard density arguments (the Stone-Weierstrass Theorem C.2.3, combined with density of $C_c(\sigma(A), \mathbb{C}) = C(\sigma(A), \mathbb{C})$ in $L^2(\sigma(A), \mu_g)$, see [79, Theorem 3.14]), one has that the space $\mathbb{P}_{\sigma(A)}[t]$ of complex valued polynomials on $\sigma(A)$ is *dense* in $L^2(\sigma(A), \mu_g)$.

Let $p \in \mathbb{P}_{\sigma(A)}[t]$ be an approximant of a given $h \in L^2(\sigma(A), \mu_g)$ such that

$$\|h - p\|_{L^2(\sigma(A),\mu_g)} < \varepsilon$$

for arbitrary $\varepsilon > 0$. Then,

$$\begin{aligned} \|h(A)g - p(A)g\|_{\mathcal{H}}^2 &= \int_{\sigma(A)} |h(t) - p(t)|^2 \, \mathrm{d}\mu_g\,(t) \\ &= \|h - p\|^2_{L^2(\sigma(A),\mu_g)} < \varepsilon^2\,, \end{aligned}$$

showing that $h(A)g$ is arbitrarily close, in L^2, to an element $p(A)g \in \mathcal{K}\,(A,\, g)$.

$$f = A^{-1}g = \int_{\sigma(A)} h(t)\, \mathrm{d}\mathbf{E}\,(t)\, g = h(A)g$$

with $h(t) = \frac{1}{t}$. Since

$$\|h\|^2_{L^2} = \int_{\sigma(A)} \frac{1}{t^2}\, \mathrm{d}\mu_g\,(t) = \|f\|^2_{\mathcal{H}} < \infty\,,$$

then by the first part of the theorem, one concludes that $f = h(A)g \in \overline{\mathcal{K}\,(A,\, g)}$. □

The map

$$\begin{aligned} \mathcal{K}\,(A,\, g) &\xrightarrow{T} L^2(\sigma(A), \mu_g) \\ p(A)g &\mapsto p\,, \end{aligned}$$

is an *isometry* because $\|p(A)g\|^2_{\mathcal{H}} = \int_{\sigma(A)} |p(t)|^2 \, \mathrm{d}\mu_g\,(t) = \|p\|^2_{L^2(\sigma(A),\mu_g)}$. Hence by density it lifts ([78, Chapter 7, Proposition 11]) to a unitary map

$$\overline{\mathcal{K}\,(A,\, g)} \xrightarrow{\cong} L^2(\sigma(A), \mu_g)\,.$$

Notice that on the one hand Proposition 4.4.4 actually proves a more general result than Krylov solvability, however, unlike Proposition 4.4.3, it does not highlight the implication A *is* $\mathcal{K}(A,\, g)$*-reduced* $\Rightarrow$ *Krylov solvability.*

The Krylov subspace gives permits one to construct a plethora of operator functions applied to the vector g.

In the proof of Proposition 4.4.3 the fact that A is $\mathcal{K}(A,\, g)$-reduced was only used to show that $A(\mathbb{1} - P_{\mathcal{K}})f \in A\mathcal{K}(A,\, g)^{\perp}$ must belong to $\mathcal{K}(A,\, g)^{\perp}$ and thus the vanishing of $A(\mathbb{1} - P_{\mathcal{K}})f$. The same argument follows merely assuming that the Krylov intersection $\mathscr{I}_{\mathcal{K}}(A,\, g)$ is trivial. For bounded bijections, triviality of $\mathscr{I}_{\mathcal{K}}(A,\, g)$ becomes necessary.

Proposition 4.4.5. *Let A be a bounded and injective operator on a Hilbert space $\mathcal{H}$, and let $f \in \mathcal{H}$ be a solution to $Af = g$, given $g \in \mathrm{ran}A$.*

(i) If $\mathscr{I}_{\mathcal{K}}(A,\, g) = \{0\}$, then $f \in \overline{\mathcal{K}(A,\, g)}$.

(ii) Assume further that A is invertible with everywhere defined, bounded inverse on $\mathcal{H}$. Then $f \in \overline{\mathcal{K}(A,\, g)}$ if and only if $\mathscr{I}_{\mathcal{K}}(A,\, g) = \{0\}$.

Proof. Part (i) and the 'if' implication of part (ii) follow from the comments just before the statement of the proposition. Conversely, if $A^{-1} \in \mathscr{B}(\mathcal{H})$ and $f \in \overline{\mathcal{K}(A,\, g)}$, then $A\overline{\mathcal{K}(A,\, g)}$ is dense in $\overline{\mathcal{K}(A,\, g)}$ (Proposition 4.4.2). Take $z \in \mathscr{I}_{\mathcal{K}}(A,\, g)$, and say $z = Aw$ for some unique $w \in \mathcal{K}(A,\, g)^{\perp}$. Based on the density above, let $(Ax_n)_{n\in\mathbb{N}}$ be a sequence in $A\overline{\mathcal{K}(A,\, g)}$ of approximants of z for some x_n's in $\overline{\mathcal{K}(A,\, g)}$. From $Ax_n \to z = Aw$ and $\|A^{-1}\|_{\mathrm{op}} < \infty$ one has $x_n \to w$ as $n \to \infty$. Since $x_n \perp w$, then

$$0 = \lim_{n\to\infty} \|x_n - w\|_{\mathcal{H}}^2 = \lim_{n\to\infty} (\|x_n\|_{\mathcal{H}}^2 + \|w\|_{\mathcal{H}}^2) = 2\,\|w\|_{\mathcal{H}}^2\,,$$

so that clearly $w = 0$ and hence $z = 0$. □

The results of Propositions 4.4.2 part (ii) and 4.4.5 part (ii) are equivalent conditions to the Krylov solvability of the well-defined (4.1). Proposition 4.4.5(ii) shows that under the conditions of bounded bijectivity of A *Krylov solvability is tantamount as the triviality of the Krylov intersection.*

This result clearly covers also the particular case when A is bounded, self-adjoint, and positive, consistently with the same conclusion obtained by Nemirovskiy and Polyak [64] through an independent analysis of the error convergence of the conjugate-gradient algorithm (see Chapter 6 for the details).

Using the *general* functional calculus for bounded operators together with some results from approximation theory (see Appendix B, Section B.3 and Appendix C, Section C.2), the following results show that a very wide class of problems are Krylov solvable. Specifically, it gives conditions under which one may approximate the inverse of an injective operator in the *operator norm* using polynomial sequences of the operator A.

For convenience, an operator $A \in \mathscr{B}(\mathcal{H})$ is said to be in *class-$\mathscr{K}$* if

(i) $0 \in \rho(A)$,

(ii) there exists some open $\mathcal{W} \subset \mathbb{C}$ such that $\sigma(A) \subset \mathcal{W}$ with $\overline{\mathcal{W}}$ compact, and in addition,

(iii) $0 \notin \overline{\mathcal{W}}$, and $\mathbb{C}^* \setminus \mathcal{W}$ is connected,

where $\mathbb{C}^*$ denotes the single point compactification of $\mathbb{C}$.

Theorem 4.4.6. *Let $A \in \mathscr{B}(\mathcal{H})$ on a Hilbert space $\mathcal{H}$ be a class-$\mathscr{K}$ operator as described above. Then one has that there exists some polynomial sequence $(p_n)_{n\in\mathbb{N}}$ such that $\|p_n(A) - A^{-1}\|_{\mathrm{op}} \to 0$ as $n \to \infty$.*

Proof. Let $\mathcal{U} \subset \mathbb{C}$ be an open set containing $\overline{\mathcal{W}}$ with $0 \notin \mathcal{U}$, so $\overline{\mathcal{W}} \subset \mathcal{U} \subset \mathbb{C}$. Then $z \mapsto z^{-1}$ is holomorphic in $\mathcal{U}$, and as such, there exist polynomials in z, $(p_n)_{n\in\mathbb{N}}$ on $\mathbb{C}$ such that

$$\left\| z^{-1} - p_n(z) \right\|_{L^\infty(\overline{\mathcal{W}})} \xrightarrow{n\to\infty} 0$$

because of Theorem C.2.5.

On the other hand, there exists a closed curve $\Gamma \subset \mathcal{W} \setminus \sigma(A)$ such that

$$\begin{aligned} \frac{1}{z} &= \frac{1}{2\pi i} \int_\Gamma \frac{1}{\zeta} (\zeta - z)^{-1} \, \mathrm{d}\zeta \\ p_n(z) &= \frac{1}{2\pi i} \int_\Gamma p_n(\zeta) (\zeta - z)^{-1} \, \mathrm{d}\zeta \, . \end{aligned}$$

because of [79, Theorem 13.5], from which also

$$\begin{aligned} A^{-1} &= \frac{1}{2\pi i}\int_\Gamma \zeta^{-1}\mathcal{R}\left(A,\,\zeta\right)\,\mathrm{d}\zeta \\ p_n(A) &= \frac{1}{2\pi i}\int_\Gamma p_n(\zeta)\mathcal{R}\left(A,\,\zeta\right)\,\mathrm{d}\zeta\,, \end{aligned}$$

as an application of Theorem B.3.2. Then the claim follows because

$$\begin{aligned} \left\|A^{-1}-p_n(A)\right\|_{\mathrm{op}} &= \left\|\frac{1}{2\pi i}\int_\Gamma\left(\frac{1}{\zeta}-p_n(\zeta)\right)\mathcal{R}\left(A,\,\zeta\right)\,\mathrm{d}\zeta\right\|_{\mathrm{op}} \\ &\leq \left\|z^{-1}-p_n(z)\right\|_{L^\infty(\overline{\mathcal{W}})}\left\|\frac{1}{2\pi i}\int_\Gamma\mathcal{R}\left(A,\,\zeta\right)\,\mathrm{d}\zeta\right\|_{\mathrm{op}}\,, \end{aligned}$$

and

$$\mathbb{1} = \frac{1}{2\pi i}\int_\Gamma\mathcal{R}\left(A,\,\zeta\right)\,\mathrm{d}\zeta\,.$$

□

Corollary 4.4.7. *Let $A\in\mathscr{B}(\mathcal{H})$ be a class-$\mathscr{K}$ operator as in Theorem 4.4.6. Consider the linear inverse problem $Af=g$ where $g\in\mathcal{H}$ and $f\in\mathcal{H}$ is a solution. Then f is a Krylov solution, i.e., $f\in\overline{\mathcal{K}\left(A,\,g\right)}$.*

Proof. By Theorem 4.4.6, one has the existence of the polynomial sequence $(p_n)_{n\in\mathbb{N}}$ that guarantees $\|p_n(A)-A^{-1}\|_{\mathrm{op}}\to 0$ as $n\to\infty$. Applying the vector g, it is immediate that $p_n(A)g\in\mathcal{K}\left(A,\,g\right)$, and one has $\|p_n(A)g-A^{-1}g\|_{\mathcal{H}}\to 0$ from which the result follows. □

An instance of a bounded operator that satisfies the conditions of Corollary 4.4.7 would be a sectorial operator with 0 outside its numerical range.

Remark 4.4.8. The unitary right shift operator on $\ell^2(\mathbb{Z})$ provides an interesting example where Corollary 4.4.7 clearly *cannot* be used, even though 0 is in the resolvent set. It is known that the singular values of unitary operators form the unit disc $\{z\in\mathbb{C};\ |z|=1\}$ in the complex plane. Therefore, any open set $\mathcal{W}$ as described in Theorem 4.4.6 is *impossible* to construct such that $\mathbb{C}^*\setminus\overline{\mathcal{W}}$ is connected and $0\notin\overline{\mathcal{W}}$. The approximation result Theorem C.2.5

may not hold. Furthermore by analogy to Example 4.4.1 (v), the unweighted right shift operator on $\ell^2(\mathbb{Z})$ is in general *not* Krylov solvable.

4.4.3 Krylov reducibility and Krylov solvability

At this point the relation between $\mathcal{K}(A,\, g)$-reducibility of A and Krylov solvability is discussed further. From Proposition 4.4.3 the former clearly implies the latter. There is more to Krylov reducibility, as for example, the following remark illustrates that the relation between the $\mathcal{K}(A,\, g)$-reducibility of A and Krylov solvability is also *equivalent* for the class of unitary operators.

Remark 4.4.9. For *unitary* operators $U : \mathcal{H} \to \mathcal{H}$, the Krylov solvability is *equivalent* to U being $\mathcal{K}(U,\, g)$-reduced. The fact that U being $\mathcal{K}(U,\, g)$-reduced implies solvability is an immediate property from Proposition 4.4.3. Conversely, as $f = U^*g$ is the solution to the linear inverse problem $Uf = g$ for some $g \in \mathcal{H}$, then the assumption that $f \in \overline{\mathcal{K}(U,\, g)}$ implies that $U^*g \in \overline{\mathcal{K}(U,\, g)}$, which by Proposition 4.3.7 is identical to the fact that U is $\mathcal{K}(U,\, g)$-reduced.

Example 4.3.9 made clear that there are cases where one has Krylov-solvability of the well-defined linear inverse problem, however one fails to have Krylov reducibility. In fact, the operator A_θ is *not* normal, and so one may naturally ask the question: is there relationship between Krylov solvability and Krylov reducibility for normal operators? The following statements reveal that a well-defined linear inverse problem with a normal operator A may indeed be Krylov solvable, but A is *not* $\mathcal{K}(A,\, g)$-reduced. First, the following feature of L^2 convergence of holomorphic functions is needed.

Theorem 4.4.10. *Let $\mathcal{U} \subset \mathbb{C}$ be an open subset of the complex plane, and $H(\mathcal{U})$ the set of all holomorphic functions on $\mathcal{U}$. Then the space $H(\mathcal{U}) \cap L^2(\mathcal{U})$ is closed in $L^2(\mathcal{U})$ in the $\|\cdot\|_{L^2}$ norm. In particular, any convergent sequence $(f_n)_{n\in\mathbb{N}} \subset H(\mathcal{U}) \cap L^2(\mathcal{U})$ converges uniformly on any compact subset of $\mathcal{U}$.*

Proof. Consider a sequence of convergent holomorphic functions $(f_n)_{n\in\mathbb{N}} \subset H(\mathcal{U}) \cap L^2(\mathcal{U})$, so that $f_n \xrightarrow{\|\cdot\|_{L^2}} f \in L^2(\mathcal{U})$. Let $K \subset \mathcal{U}$ be a compact set, and define δ' as the distance

$$\delta' := \mathrm{dist}(K, \mathbb{C} \setminus \mathcal{U})\,.$$

Let $z_0 \in K$ and $0 < r < \delta \leq \delta'$ such that $\delta < \infty$; and clearly $z_0 + r\exp(i\theta) \in \mathcal{U}$ for all $\theta \in \mathbb{R}$. As $f_n \to f \in L^2(\mathcal{U})$ the sequence is Cauchy, let $\varepsilon > 0$ so that $\exists\, N \in \mathbb{N}$ such that $\forall\, n, m \geq N$ one has $\|f_n - f_m\|_{L^2} < \epsilon$. Using the mean value property [79, Chapter 11] and the Cauchy-Schwartz inequality one has

$$\begin{aligned}
&|f_n(z_0) - f_m(z_0)| = \left|\frac{1}{2\pi}\int_0^{2\pi} f_n\left(z_0 + r\exp(i\theta)\right) - f_m\left(z_0 + r\exp(i\theta)\right)\,\mathrm{d}\theta\right| \\
&\quad\leq \frac{1}{2\pi}\int_0^{2\pi} \left|f_n\left(z_0 + r\exp(i\theta)\right) - f_m\left(z_0 + r\exp(i\theta)\right)\right|\,\mathrm{d}\theta \\
&\quad\leq \frac{1}{2\pi}\left(\int_0^{2\pi} 1\,\mathrm{d}\theta\right)^{\frac{1}{2}}\left(\int_0^{2\pi} \left|f_n\left(z_0 + r\exp(i\theta)\right) - f_m\left(z_0 + r\exp(i\theta)\right)\right|^2\mathrm{d}\theta\right)^{\frac{1}{2}} \\
&\quad= \left(\frac{1}{2\pi}\int_0^{2\pi} \left|f_n\left(z_0 + r\exp(i\theta)\right) - f_m\left(z_0 + r\exp(i\theta)\right)\right|^2\mathrm{d}\theta\right)^{\frac{1}{2}}\,.
\end{aligned}$$

Taking the square and integrating further in the variable $r\,\mathrm{d}r$ over the interval $(\delta/2, \delta)$,

$$\begin{aligned}
\frac{3\delta^2}{4}|f_n(z_0) - f_m(z_0)|^2 &= \int_{\delta/2}^{\delta} |f_n(z_0) - f_m(z_0)|^2 r\,\mathrm{d}r \\
&\leq \frac{1}{2\pi}\int_{\delta/2}^{\delta}\int_0^{2\pi} \left|f_n\left(z_0 + r\exp(i\theta)\right) - f_m\left(z_0 + r\exp(i\theta)\right)\right|^2 r\,\mathrm{d}r\,\mathrm{d}\theta\,.
\end{aligned}$$

As $z_0 + r\exp(i\theta) \in \mathcal{U}$ for all θ and for all r over the integration interval,

$$\begin{aligned}
\int_{\delta/2}^{\delta}\int_0^{2\pi} &\left|f_n\left(z_0 + r\exp(i\theta)\right) - f_m\left(z_0 + r\exp(i\theta)\right)\right|^2 r\,\mathrm{d}r\,\mathrm{d}\theta \\
&\leq \|f_n - f_m\|^2_{L^2(\mathcal{U})} < \varepsilon^2\,.
\end{aligned}$$

Putting everything together

$$\frac{3\delta^2}{4}|f_n(z_0) - f_m(z_0)|^2 < \varepsilon^2\,,$$

so that the sequence $(f_n)_{n\in\mathbb{N}}$ is also Cauchy in the uniform convergence topology on the compact set K. The continuous functions are a complete space in the uniform convergence topology [59]. So there exists some $\tilde{f} : K \to \mathbb{C}$, a continuous function on the compact set K, such that the restriction of the functions f_n to K, i.e. $f_n|_K$, converges uniformly to $\tilde{f}$. It remains to be shown that in fact $\tilde{f} = f|_K$. Clearly on the compact set K, one has $\|f_n - f\|_{L^2(K)} \to 0$ by assumption, and also $\|f_n - \tilde{f}\|_{L^2(K)} \to 0$ as $n \to \infty$ by the above. Then by the triangle inequality, $\|\tilde{f} - f\|_{L^2(K)} = 0$, so that $\tilde{f}$ and f are the same almost everywhere on K. So $f = \tilde{f}$ everywhere on K. So $f_n \to f$ uniformly on compact sets $K \subset \mathcal{U}$. Therefore, by [79, Theorem 10.28] one immediately sees that $f \in H(\mathcal{U})$. □

The example below finally answers the question about the connection between Krylov solvability and Krylov reducibility for normal operators. In general, the former does *not* imply the latter.

Example 4.4.11. The multiplication operator M_z of Example 4.4.1(ii) is used here. From this, it is known that the problem $M_z f = g$ for $g \in L^2(\Omega)$ is *always* Krylov solvable. However, there is a choice of vector g that one can make such that M_z is *not* $\mathcal{K}(M_z,\, g)$-reduced.

Indeed, let $g \in L^2(\Omega)$ be such that $0 < \varepsilon \leq |g(z)| \leq \varepsilon' < \infty$ on Ω. The Krylov space of this problem is

$$\overline{\mathcal{K}(M_z,\, g)} = \left\{\phi g;\ \phi \in \overline{\mathbb{P}_\Omega[z]}^{\|\cdot\|_2}\right\}, \tag{4.9}$$

(where $\overline{E}^{\|\cdot\|_2}$ denotes the closure of a set E in the $L^2(\Omega)$ topology) and $\overline{\mathbb{P}_\Omega[z]}^{\|\cdot\|_2}$ is contained in the holomorphic functions on Ω (Theorem 4.4.10). Surely, consider the space $\mathcal{K}(M_z,\, g)$,

$$\mathcal{K}(M_z,\, g) = \{pg;\ p \in \mathbb{P}_\Omega[z]\}.$$

Let $w \in \overline{\mathcal{K}(M_z,\, g)}$, so that there exists a sequence $(w_n)_{n\in\mathbb{N}} \to w$ where $w_n = p_n g$ is in $\mathcal{K}(M_z,\, g)$. The sequence $(p_n g)_{n\in\mathbb{N}}$ is Cauchy in $L^2(\Omega)$ and one has that $(p_n)_{n\in\mathbb{N}}$ is also Cauchy. Indeed,

$$\|p_n - p_m\|_2 = \left\| \frac{1}{g}(gp_n - gp_m) \right\|_2 \leq \frac{1}{\varepsilon} \|gp_n - gp_m\|_2 \xrightarrow{n,m\to\infty} 0\,,$$

so that $p_n \xrightarrow{\|\cdot\|_2} \phi \in L^2(\Omega) \cap \overline{\mathbb{P}_\Omega[z]}^{\|\cdot\|_2}$. The uniqueness of the limit guarantees that $w = \phi g$.

To show the reverse inclusion, let $w = \phi g$ for some $\phi \in \overline{\mathbb{P}_\Omega[z]}^{\|\cdot\|_2}$. Therefore, there exists a polynomial sequence $(p_n)_{n\in\mathbb{N}}$ such that $p_n \xrightarrow{\|\cdot\|_2} \phi$, and $p_n g \to \phi g$ in $L^2(\Omega)$ as $\|p_n g - g\phi\|_2 \leq \varepsilon' \|p_n - \phi\|_2 \to 0$ as $n \to \infty$. Then it is clear that $w \in \overline{\mathcal{K}(M_z,\, g)}$.

Finally, one now sees that the problem $M_z f = g$ is *not* $\mathcal{K}(M_z,\, g)$-reduced. Indeed, consider that the adjoint operator of M_z is the mapping $f \to \overline{z}f$, i.e., $M_z^* = M_{\overline{z}}$. By the Cauchy-Riemann relations it is obvious that $\overline{z}$ is *not* holomorphic anywhere on $\mathbb{C}$, let alone on Ω. Using Proposition 4.3.7 one may proceed by showing that it is impossible for $M_z^* g \in \overline{\mathcal{K}(M_z,\, g)}$. Proceeding by contradiction, assume that $\overline{z}g \in \overline{\mathcal{K}(M_z,\, g)}$ which implies that $\overline{z} \in \overline{\mathbb{P}_\Omega[z]}^{\|\cdot\|_2}$. However, as $\overline{\mathbb{P}_\Omega[z]}^{\|\cdot\|_2} \subset H(\Omega) \cap L^2(\Omega)$, one comes to the contradiction that $\overline{z}$ is holomorphic.

4.4.4 Krylov solutions in the lack of injectivity

Finally under consideration is the scenario where one has a solvable linear inverse problem (i.e. $g \in \mathrm{ran}A$) with A *not* injective. Krylov reducibility still guarantees the existence of Krylov solutions, and under certain assumptions on the kernel of the operator, one may even show the *uniqueness* of the solution in the Krylov subspace. The following proposition is the counterpart to Proposition 4.4.3 under (possible) lack of injectivity of A.

Proposition 4.4.12. *Let A be a bounded linear operator on a Hilbert space $\mathcal{H}$, and let $g \in \mathrm{ran}A$. If A is $\mathcal{K}(A,\, g)$-reduced, then there exists a Krylov solution to the problem $Af = g$. For example, if $f_\circ \in \mathcal{H}$ satisfies $Af_\circ = g$ and $P_\mathcal{K}$ is the orthogonal projection onto $\overline{\mathcal{K}(A,\, g)}$, then $f := P_\mathcal{K} f_\circ$ is a Krylov solution.*

Proof. Let $f_\circ \in \mathcal{H}$ be any vector that satisfies $Af_\circ = g$. From the same arguments in the proof of Proposition 4.4.3 one has $A(\mathbb{1} - P_{\mathcal{K}}f_\circ) = 0$. Thus $AP_{\mathcal{K}}f_\circ = g$, i.e. $f := P_{\mathcal{K}}f_\circ$ is a Krylov solution. □

Although general bounded linear inverse problems may exhibit more than one solution, some of which might not be in the Krylov space (see Example 4.4.1), for a fairly general class of problems one may prove that the Krylov solution, when it exists, is *unique.*

Proposition 4.4.13. *Let A be a bounded linear operator on Hilbert space $\mathcal{H}$ and let $Af = g$ be the associated linear inverse problem, given $g \in \mathrm{ran}A$. If $\ker A \subset \ker A^*$, then there exists at most one solution $f \in \overline{\mathcal{K}(A, g)}$. In particular, the same conclusion holds for bounded normal operators.*

Proof. If $f_1, f_2 \in \overline{\mathcal{K}(A, g)}$ and $Af_1 = g = Af_2$, then $f_1 - f_2 \in \ker A \cap \overline{\mathcal{K}(A, g)}$. By hypothesis $\ker A \subset \ker A^*$, and rather obviously $\overline{\mathcal{K}(A, g)} \subset \overline{\mathrm{ran}A}$. As such, $f_1 - f_2 \in \ker A^* \cap \overline{\mathrm{ran}A}$. But $\ker A \cap \overline{\mathrm{ran}A} = \{0\}$, whence $f_1 = f_2$. The second statement follows from the fact that for a normal operator one has $\ker A = \ker A^*$. □

The above proposition is similar to some comments made in [30, 11, 32] about Krylov solutions to singular systems in finite-dimensions.

A consequence of Proposition 4.4.13 is the following corollary for self-adjoint operators.

Corollary 4.4.14. *If $A \in \mathscr{B}(\mathcal{H})$ is self-adjoint, then the linear inverse problem $Af = g$ with $g \in \mathrm{ran}A$ admits a unique Krylov solution.*

Proof. The operator A is $\mathcal{K}(A, g)$-reduced (Remark 4.3.5), and hence the linear inverse problem admits a Krylov solution (Proposition 4.4.12). Such a solution is then unique, owing to Proposition 4.4.13. □

4.5 Numerical Tests and Examples

This Section is aimed at providing some numerical tests that, despite their simplicity, reveal several features discussed in the previous sections. These tests are intended to be of pedagogical value for the theoretical points made rather than an in-depth numerical study of any particular algorithm. Throughout, the GMRES algorithm of [86] is used due to its generality of being applied to general operator classes.

The focus on the behaviour of the convergence of the residual and error terms occurs under the following circumstances:

1. when the solution f to the injective problem $Af = g$ for $g \in \mathrm{ran}A$ is or is not a Krylov solution,

2. when the linear operator is not injective (well-defined vs ill-defined problems).

To begin with, the various problems and methods are outlined. Then the situation of Krylov vs non-Krylov solutions is examined for an injective operator. Lastly the case of a lack of injectivity is explored.

4.5.1 Four inverse linear problems

The 'baseline' case considered, where the solution is known a-priori to be a Krylov solution, is a compact, injective, self-adjoint multiplication operator on $\ell^2(\mathbb{N})$ (see Appendix A),

$$M = \sum_{n=1}^{\infty} \sigma_n \left| e_n \right\rangle \left\langle e_n \right| \,, \quad \sigma_n = (5n)^{-1} \,. \tag{4.10}$$

In comparison to M, tests of the non-injective version,

$$\widetilde{M} = \sum_{n=1}^{\infty} \widetilde{\sigma}_n \left| e_n \right\rangle \left\langle e_n \right| \,, \quad \widetilde{\sigma}_n = \begin{cases} 0 & \text{if } n = \{3,6,9\} \\ \sigma_n & \text{otherwise} \,. \end{cases} \,, \tag{4.11}$$

are also presented, and so too are results for the weighted right shift (Appendix A)

$$\mathcal{R} = \sum_{n=1}^{\infty} \sigma_n \left| e_{n+1} \right\rangle \left\langle e_n \right| \,, \tag{4.12}$$

where the σ_n's are the same as in (4.10). The linear inverse problems $Mf = g$, $\widetilde{M}f = g$ and $\mathcal{R}f = g$ are investigated with a datum g generated by the a-priori chosen solution

$$f = \sum_{n\in\mathbb{N}} f_n e_n\,, \quad f_n = \begin{cases} n^{-1} & \text{if } n \leqslant 250 \\ 0 & \text{otherwise}\,. \end{cases} \tag{4.13}$$

One has that

$$\|f\|_{\ell^2} = \sqrt{\frac{\pi^2}{6} - \psi^{(1)}(251)} \approx 1.28099\,, \tag{4.14}$$

where $\psi^{(k)}$ is the polygamma function of order k (see [1, Section 6.4]).

The final problem considered here is the linear inverse problem $Vf = g$, where V is the Volterra operator on $L^2[0,1]$ (see Examples 4.2.2 and 4.4.1) and $g(x) = \frac{1}{2}x^2$, so that the problem has unique solution

$$f(x) = x\,, \quad \|f\|_{L^2} = \frac{1}{\sqrt{3}} \approx 0.5774\,. \tag{4.15}$$

The inverse linear problems associated to M and $\widetilde{M}$ are Krylov solvable (Corollary 4.4.14) as well as V (Example 4.4.1 (vii)). The inverse problem associated to $\mathcal{R}$ is *not* Krylov solvable as the space $\mathcal{K}(\mathcal{R},\, g)^\perp$ contains the canonical vector e_1. For ease of discussion, depending on the context, $\mathcal{H}$ and A will respectively denote the Hilbert space ($\ell^2(\mathbb{N})$ or $L^2[0,1]$) and operator (M, $\widetilde{M}$, $\mathcal{R}$, or V).

The numerical tests presented herein on $\ell^2(\mathbb{N})$ (i.e., for M, $\widetilde{M}$ and $\mathcal{R}$) generate the spanning vectors $g, Ag, A^2g, \dots$ of $\mathcal{K}(A,\, g)$ represented in the canonical basis up to order $N_{\max} = 500$. When $A = V$, the spanning vectors are constructed up to $N_{\max} = 175$ represented using a Legendre polynomial basis on $[0,1]$. These values represent a practical choice of 'infinite' dimension for $\mathcal{K}(A,\, g)$.

When $A = M, \widetilde{M}, \mathcal{R}$, the space of entries allocated for each of the considered vectors, $f, g, Ag, A^2g, \dots$, is 2500 entries with respect to the canonical basis of $\ell^2(\mathbb{N})$. Again, this is to represent a practical choice of 'infinite' dimension of the ambient Hilbert space $\mathcal{H}$. To illustrate this point, in particular, if one considers the repeated application of $\mathcal{R}$ up to 500 times, the vectors $\mathcal{R}^k g$ have non-trivial entries up to order $251 + 500 = 751$. This is because, by construction, the last non-zero entries of f and g are the respective components e_{250} and e_{251}. Also, by repeated application of M and $\widetilde{M}$, the vectors $M^k g$ and $\widetilde{M}^k g$ have the component e_{250} as the last non-trivial entry. All these limits are well below the choice of the 'infinite' dimensional threshold of 2500 entries for $\mathcal{H}$.

On the other hand, for $A = V$, the practical choice of 'infinite' dimension for $\mathcal{H}$ was 250. In this case, it is expected that the numerical tests contain no significant numerical errors in the computation with respect to the Legendre basis polynomials, as the $L^2[0,1]$ norm of each basis has less than 2% error compared to the exact unit value.

During the course of running the numerical tests, from each collection $\{g, Ag, \dots, A^{N-1}g\}$ an orthonormal basis of the N-dimensional space $\mathcal{K}_N(A,\, g)$ is obtained. Of course, $N \leq N_{\max}$, and the 'infinite-dimensional' problem $Af = g$ is truncated to a finite N-dimensional problem using the GMRES algorithm. This is all very much in the same spirit as Chapter 3.

Therefore, $\widehat{f^{(N)}} \in \mathcal{H}$ denotes the solution at the N-th step, or the N-th iterate, from the GMRES algorithm. Both the infinite-dimensional error $\mathscr{E}_N$ and infinite-dimensional residual $\mathfrak{R}_N$ (see Definition 3.2.5) were analysed as the two natural indicators of the convergence of the iterates to a solution $f \in \mathcal{H}$ as '$N \to \infty$'.

4.5.2 Krylov vs non-Krylov solutions

Figure 4.1 illustrates the behaviours of the norms of $\mathscr{E}_N$ and $\mathfrak{R}_N$, as well as the approximated solution $\widehat{f^{(N)}}$, all as a function of the iteration number N. The numerical evidence shows that

(i) The error norm of the baseline case and the Volterra case appear to vanish with N, as too does the residual norm. This is consistent with the obvious property that $\|\mathfrak{R}_N\|_{\mathcal{H}} \leq \|A\|_{\mathrm{op}} \|\mathscr{E}_N\|_{\mathcal{H}}$. Also, $\|\widehat{f^{(N)}}\|_{\mathcal{H}}$ stays uniformly bounded and tends to the prescribed theoretical value from

(4.14) and (4.15).

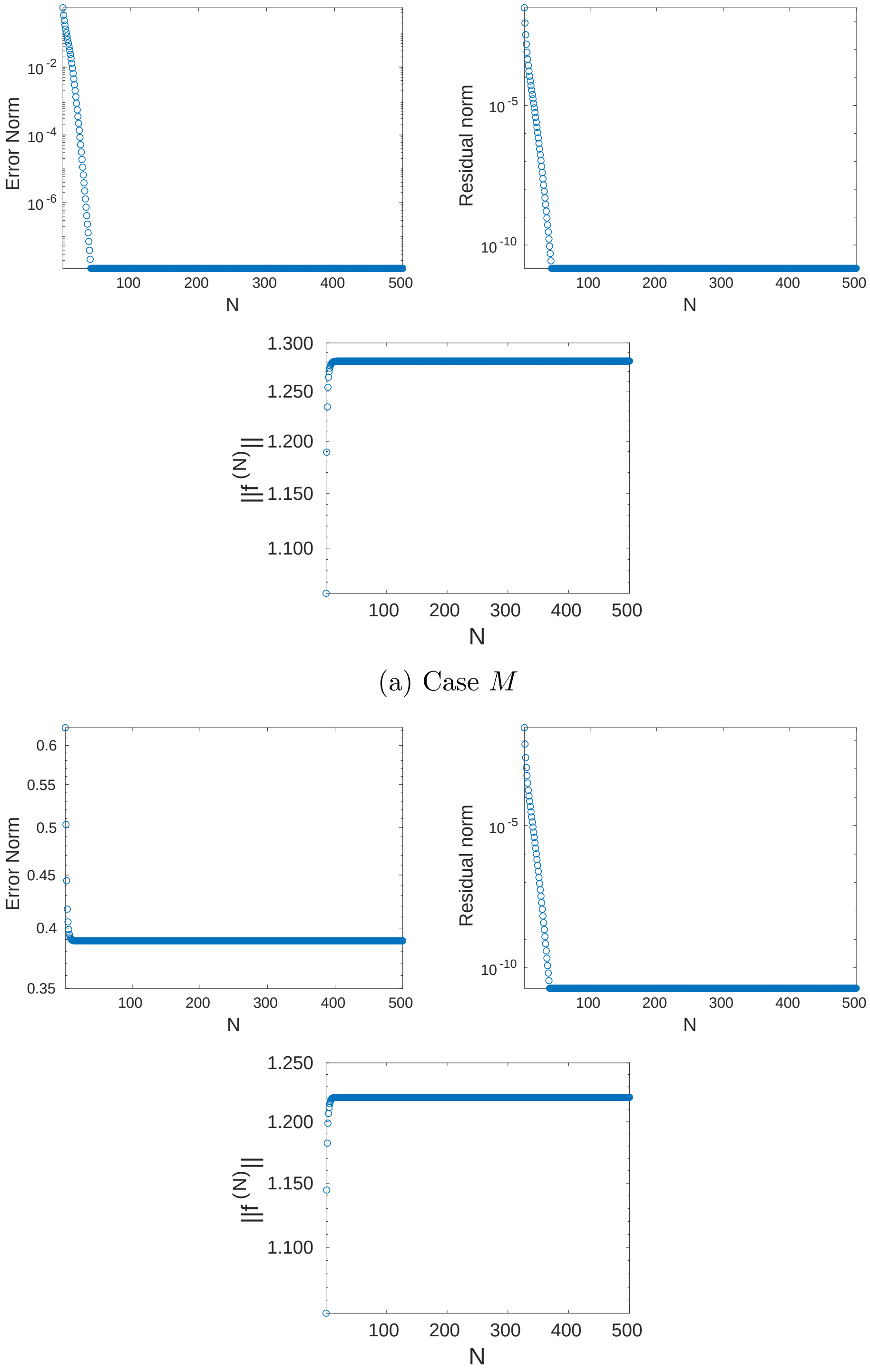
(ii) The error norm of the forward shift remains of order one indicating a *lack of convergence in the strong topology* to the solution f, regardless of the truncation size N. Analogous behaviour is also seen in the residual. Furthermore, $\|\widehat{f^{(N)}}\|_{\mathcal{H}}$ remains uniformly bounded, but it attains an asymptotic value that is strictly smaller than the theoretical value (4.14).

At this point, the asymptotics of $\|\mathscr{E}_N\|_{\mathcal{H}} \to 1$ and $\|\mathfrak{R}_N\|_{\mathcal{H}} \to 0.2$ for the problem where $A = \mathcal{R}$ are described. Since $\widehat{f^{(N)}} \in \mathcal{K}(\mathcal{R}, g)$ and since the latter subspace only contains vectors with zero component along the direction e_1, the error vector $\mathscr{E}_N = f - \widehat{f^{(N)}}$ tends to asymptotically approach the vector e_1, as this gives the first component of $f = (1, \frac{1}{2}, \frac{1}{3}, \dots)$, and explains why $\|\mathscr{E}_N\|_{\mathcal{H}} \to 1$.

Similarly, as by construction $g = (0, \frac{1}{5}, \frac{1}{20}, \frac{1}{45}, \dots)$, and since the asymptotics of the error imply that each component of $\widehat{f^{(N)}}$, except for the first one, converges to the corresponding component of f, then $\widehat{f^{(N)}} \approx (0, \frac{1}{2}, \frac{1}{3}, \dots)$ for N large. Therefore $\mathfrak{R}_N \approx (0, 0, \frac{1}{20}, \frac{1}{45}, \dots)$. As such, g and $\mathcal{R}\widehat{f^{(N)}}$ differ by only the vector $\frac{1}{5}e_2$, explaining why $\|\mathfrak{R}_N\|_{\mathcal{H}} \to 0.2$.

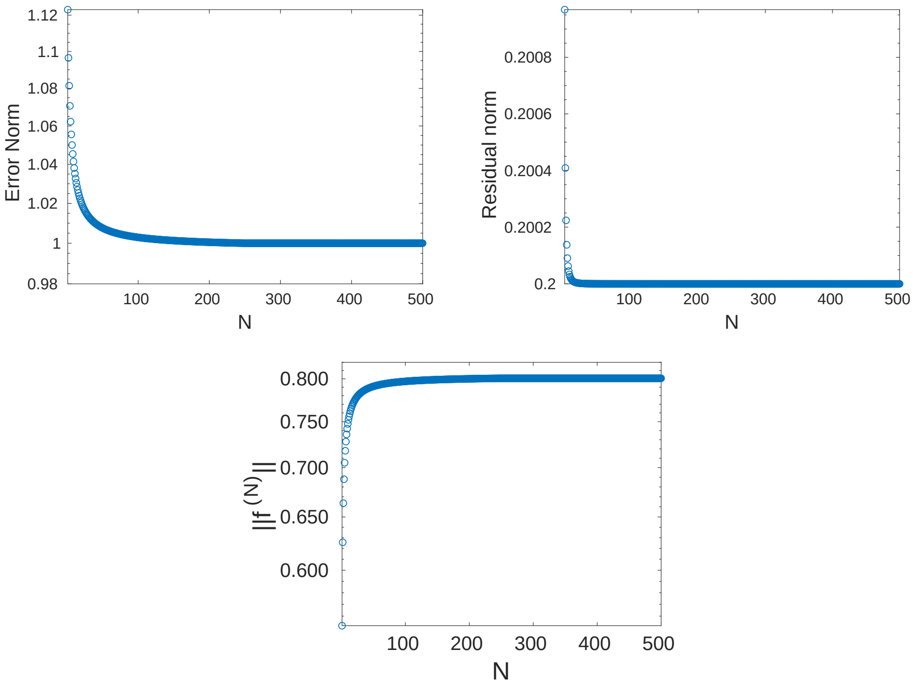
As already pointed out in Chapter 3, it is clear that the lack of vanishing of the error and residual for the problem $\mathcal{R}f = g$ do not necessarily mean that the approximations $\widehat{f^{(N)}}$ carry no information on the solution f. In fact, here the vector entries of f, apart from the first, are well approximated by $\widehat{f^{(N)}}$.

So, the Krylov solvable infinite-dimensional problems all display norm convergence in the error and residual terms. The convergence behaviour for the multiplication operator M is faster than that of the Volterra operator V. This indicates that the choice of the Krylov bases in these two cases is not equally as effective. Heuristically, it is well-known that the GMRES method may behave poorly in some circumstances [66]. However, the GMRES method applied to the self-adjoint M is mathematically equivalent to the MINRES technique, and the convergence behaviour in this case of M, a positive definite operator, has already been discussed before by [64] (see Section 2.3.1).

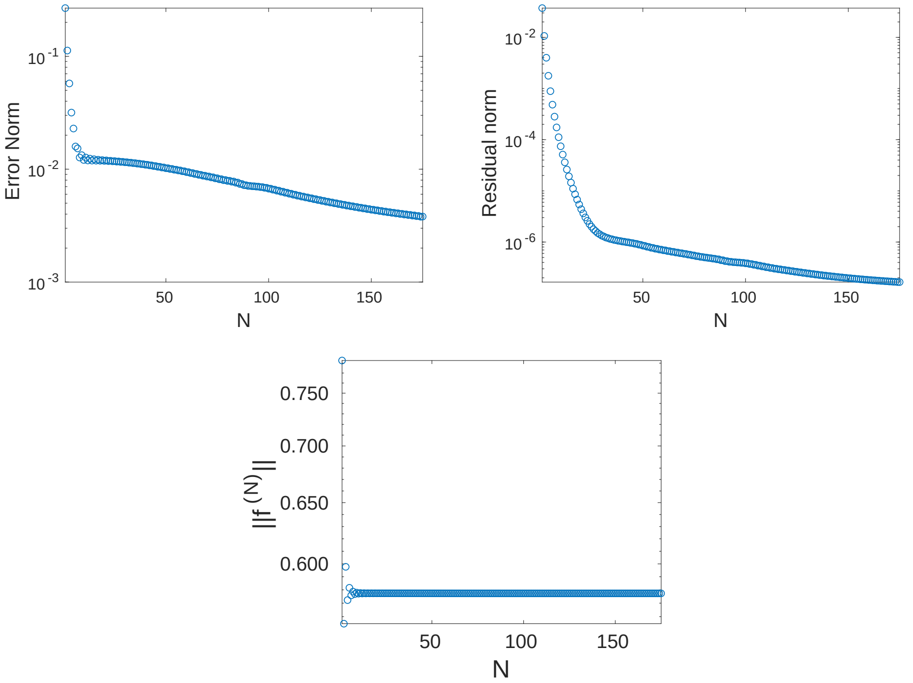


(a) Case M

(b) Case $\widetilde{M}$



(c) Case $\mathcal{R}$



(d) Case V

Figure 4.1: Error norm and residual norm as a function of iterations for the cases of the injective multiplication operator M (baseline case), the weighted right shift $\mathcal{R}$, the non-injective multiplication operator $\widetilde{M}$, and the Volterra operator V.

In contrast, the non-Krylov solvable problem $\mathcal{R}f = g$ does not exhibit norm convergence of the approximations to the solution at all. Here, the uniformity of the size of the solutions $\widehat{f^{(N)}}$ does not appear affected by the lack of Krylov solvability, and they all remain uniformly bounded.

4.5.3 Lack of injectivity

In the numerical test of the solvable inverse problem $\widetilde{M}f = g$ with $g \in \mathrm{ran}\widetilde{M}$, one has an *infinity* of solutions. Yet, in this lack of injectivity, Corollary 4.4.14 guarantees that such a problem admits a *unique* Krylov solution. Numerically it was found that (Figure 4.1)

(i) In contrast to the baseline case M, the infinite-dimensional error norm $\|\mathscr{E}_N\|_{\mathcal{H}}$ does *not* vanish with increasing truncation size N. The infinite-dimensional residual norm $\|\mathfrak{R}_N\|_{\mathcal{H}}$ on the other hand *does* display a convergence to zero, and in fact has the same behaviour as the baseline case M.

(ii) The norm of the approximated solution $\|\widehat{f^{(N)}}\|_{\mathcal{H}}$ remains uniformly bounded.

Figure 4.2 unmasks the *apparent* lack of convergence of the numerical approximants $\widehat{f^{(N)}}$ to the solution f seen by the non-vanishing error norm. One may see that the only non-zero components of the error term are precisely the vector entries corresponding to the kernel of $\widetilde{M}$. This information underscores that the GMRES algorithm has indeed found the minimal norm solution to the problem.

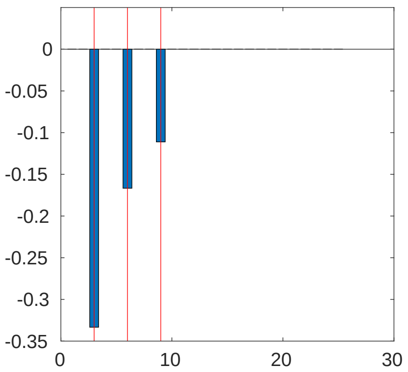


Figure 4.2: Support of the error vector (blue bars) for the non-injective problem $\widetilde{M}f = g$ at final iteration $N = 500$. The red lines mark the entry positions of the components of the kernel space of $\widetilde{M}$.

Chapter 5

Krylov Solvability for Unbounded Inverse Problems

5.1 Introduction

This Chapter is an extension of the previous work presented in Chapter 4 and is based on the work [15]. The appropriate notions of Krylov reducibility and the Krylov intersection are extended to encompass the possibility that one is dealing with an unbounded operator. Within this Chapter, the class of operators considered is broadened to the class of *closed and densely defined* operators on Hilbert space $\mathcal{H}$. Immediately domain issues come into play, however this is dealt with by making natural and sensible assumptions on the generating vector of the Krylov spaces to ensure that the standard Krylov space is indeed well defined.

Some preliminary investigations into Krylov methods in the area of unbounded linear inverse problems have been undertaken by Gilles and Townsend [35] and Olver [69]. Within both these studies however, the authors have restricted themselves to particular algorithms, such as conjugate-gradients or GMRES, for very precisely defined differential linear inverse problems. As such, there is no existing discussion of the *general* mechanisms of Krylov solvability for unbounded, closed, densely defined operators on Hilbert space.

The problem considered in this Chapter is still the linear inverse problem

$$Af = g\,, \tag{5.1}$$

except now that $A \in \mathscr{C}(\mathcal{H})$ and $\mathcal{D}(A) \subset \mathcal{H}$ is dense in $\mathcal{H}$, with $g \in \mathcal{H}$, $f \in \mathcal{H}$. Again, the usual concepts of *solvable, well-defined*, and *well-posed* apply without modification from Chapter 4 to (5.1) (see page 72). It is immediate that the adjoint A^* is closed, as $\mathcal{D}(A)$ is dense [88, Prop. 1.6].

To begin with, the following section contains the suitably generalised definition of the Krylov space. Preliminaries and definitions of *rational* Krylov subspaces will be made clear in this Section too.

Following the next section, the extensions of the major concepts regarding Krylov solvability in Chapter 4 are made, revealing that they are indeed the *intrinsic* notions of Krylov solvability at a suitably general level. However, it should be pointed out that under such general conditions, some extra (yet natural) hypotheses are required.

Towards the end of this Chapter, some aspects of rational Krylov subspaces and rational Krylov solvability are discussed within the framework of self-adjoint operators.

5.2 Definitions and comments

Immediately the very notion of a standard Krylov subspace is called into question. It may not be the case that every power of the operator A applied to some $g \in \mathcal{H}$ is defined. Therefore it is necessary to have additional assumptions to ensure that the standard Krylov subspace is well-defined.

To start with, the space of what will be called "smooth vectors" with respect to the (possibly unbounded) operator A, is defined. This additional notation assists with building up the appropriate notion of standard Krylov subspaces.

Definition 5.2.1. A vector $g \in \mathcal{H}$ is said to be *N-regular*, for a given $N \in \mathbb{N}_0$, with respect to a linear operator $A : \mathcal{H} \to \mathcal{H}$ on Hilbert space $\mathcal{H}$, if g is in the domain of all positive integer powers of A less than or equal to N. The customary notation for this space is

$$\mathcal{C}^N(A) := \bigcap_{n=0}^{N} \mathcal{D}(A^n) \subset \mathcal{H}, \tag{5.2}$$

and moreover, if $N = \sup\{n \,|\, g \in \mathcal{C}^n(A)\} < \infty$, then g is said to be *strictly* N-regular. In addition, $g \in \mathcal{H}$ is said to be a *smooth vector* with respect to A if belongs to all positive integer powers of the domain of the operator A. The customary notation for this space is

$$\mathcal{C}^\infty(A) := \bigcap_{n\in\mathbb{N}_0} \mathcal{D}(A^n) \subset \mathcal{H}. \tag{5.3}$$

Using the above notation, the standard Krylov subspaces may now be defined as follows.

Definition 5.2.2. Let $\mathcal{H}$ be a Hilbert space, and let $A : \mathcal{H} \to \mathcal{H}$ be a closed, densely defined linear operator with domain $\mathcal{D}(A) \subset \mathcal{H}$. Consider some $g \in \mathcal{C}^{N-1}(A)$ for some $N \in \mathbb{N}$. Then the N-th order Krylov subspace associated with A and g is

$$\mathcal{K}_N(A, g) := \operatorname{span}\{A^n g \,|\, n \in \mathbb{N}_0,\, n \leq N-1\}\,, \tag{5.4}$$

and further assuming that $g \in \mathcal{C}^\infty(A)$, then *the* Krylov subspace associated with A and g is

$$\mathcal{K}(A, g) := \operatorname{span}\{A^n g \,|\, n \in \mathbb{N}_0\}\,. \tag{5.5}$$

From this definition, it is obvious that $\mathcal{K}(A, g)$ remains invariant under the action of A, i.e., $A\mathcal{K}(A, g) \subset \mathcal{K}(A, g)$. This does *not* immediately carry over to the closure of $\mathcal{K}(A, g)$ because of the obvious domain issues.

Example 5.2.3. Consider the Laplacian operator $\Delta : L^2(\Omega) \to L^2(\Omega)$ where $\Omega \subset \mathbb{R}^N$ is open (and $N \in \mathbb{N}$), and $\mathcal{D}(\Delta)$ is the set of twice differentiable functions on $L^2(\Omega)$ with Dirichlet boundary conditions. As $C_c^\infty(\Omega)$, the space of smooth, compactly supported functions in Ω, is dense in $L^2(\Omega)$ [10], it is evident that Δ has dense domain.

Now, let $g \in C_c^\infty(\Omega)$. Then $g \in \mathcal{C}^\infty(\Delta)$, so that the Krylov space associated with Δ and g is appropriately defined.

Next to the notion of a standard (polynomial) Krylov subspace, a new type of Krylov subspace is introduced, namely that of *rational* Krylov subspaces. In the previous definitions, Krylov spaces are built using the positive powers of the operator A applied to a vector g, and so are known as *polynomial* or *standard* Krylov subspaces (or just 'Krylov subspace'). The concept of a rational Krylov subspace uses the class of *rational* functions of the operator A applied to a vector g to build up a vector space. Below is a general definition that will be used later.

Definition 5.2.4. Let $\mathcal{H}$ be a Hilbert space, and let $A : \mathcal{H} \to \mathcal{H}$ be a closed, densely defined linear operator with domain $\mathcal{D}(A) \subset \mathcal{H}$. Consider some sequence $\Xi \equiv (\xi_n)_{n\in\mathbb{N}} \subset \rho(A)$. Then the N-th order rational Krylov subspace associated with A, g, and Ξ is

$$\mathcal{K}_N^\Xi(A,\, g) := \operatorname{span}\{g\} \oplus \operatorname{span}\left\{\prod_{n=1}^{m}(A - \xi_n \mathbb{1})^{-1} g;\ 1 \leq m \leq N-1\right\}, \tag{5.6}$$

and *the* rational Krylov subspace associated with A, g, and Ξ is

$$\mathcal{K}^\Xi(A,\, g) := \operatorname{span}\{g\} \oplus \operatorname{span}\left\{\prod_{n=1}^{m}(A - \xi_n \mathbb{1})^{-1} g;\ m \in \mathbb{N}\right\}. \tag{5.7}$$

The concept of rational Krylov spaces was first explicitly introduced in the finite-dimensional setting by Ruhe [82] for solving the eigenvalue problem, and is now an accepted and well-studied area of numerical literature (see [26, 84, 82, 31, 80, 81, 83, 58, 94, 43, 48, 23, 7, 39]).

These spaces are particularly attractive in the study of numerical techniques of time-dependent partial differential equations [31, 58, 94, 23] as they have applications in approximating time dependent functions used in the time-stepping of the solution. However, typically these studies have focussed on the approximation of certain operator functions and their eigenvalue eigenvector pairs, rather than the Krylov solvability of the linear inverse problem.

Furthermore, these studies are mostly restricted to the finite-dimensional setting, and are also restricted to the class of bounded linear operators.

In the final section of this Chapter, some aspects of rational Krylov spaces and the linear inverse problem are considered for the class of unbounded self-adjoint operators in Hilbert space.

Some operator-theoretic notions are also needed to develop further the theory in the following sections. To begin with, the notion of a *part* of a linear operator is defined on a closed subspace $\mathcal{V} \subset \mathcal{H}$, consistent with the concept as described in [88, Definition 1.7].

Definition 5.2.5. Let $A : \mathcal{H} \to \mathcal{H}$ be a linear operator on the Hilbert space $\mathcal{H}$, and let $\mathcal{V} \subset \mathcal{H}$ be a closed subspace. Let A_0 be a domain restriction of A with $\mathcal{D}(A_0) = \mathcal{V} \cap \mathcal{D}(A)$ for some $\mathcal{V} \subset \mathcal{H}$. Then the operator A_0 is called the *part* of the operator A on $\mathcal{V}$.

The notion of the *core* of an operator is also necessary for the purposes of the proofs of the following lemmas and propositions, and may be found in many standard references on operator theory (e.g., [51, 88]).

Definition 5.2.6. Let A be a closed operator on $\mathcal{H}$. Let T be an operator that has a closed extension $\overline{T}$, i.e., it is *closable.* If $\overline{T} = A$, then $\mathcal{D}(T)$ is called a *core* of the operator A, and $\overline{G(T)} = G(A)$. Equivalently, a linear submanifold $\mathcal{D} \subset \mathcal{D}(A)$ is called a core of A if the set $\{(u, Au) \,|\, u \in \mathcal{D}\}$ is *dense* in $G(A)$.

The core of an operator is a powerful concept that will be used to ensure that *invariance* of the Krylov subspace under A still occurs at a suitably general level.

The concept of the *invariance* of a closed subspace under the action of an operator A is generalised appropriately.

Definition 5.2.7 (see Definition 1.7 [88])**.** Given a linear operator $A : \mathcal{H} \to \mathcal{H}$ on Hilbert space $\mathcal{H}$ and a closed subspace $\mathcal{V} \subset \mathcal{H}$, then one says that $\mathcal{V}$ is *invariant* under the action of A if $A(\mathcal{V} \cap \mathcal{D}(A)) \subset \mathcal{V}$.

As these general definitions have now been set, one may proceed with the development of some necessary conditions to ensure the invariance of the closed Krylov subspace $\overline{\mathcal{K}(A,\, g)}$ under the action of A.

5.3 Krylov reducibility and Krylov intersection

The fundamental concepts of Krylov reducibility and Krylov intersection are suitably generalised to the class of closed, densely defined operators on Hilbert space.

For a given closed, densely defined operator A, and a given vector $g \in \mathcal{C}^\infty(A)$, one still has the orthogonal decomposition of the Hilbert space (Chapter 4, page 75), namely

$$\mathcal{H} = \overline{\mathcal{K}(A,\, g)} \oplus \mathcal{K}(A,\, g)^\perp\ . \tag{4.4}$$

This is still referred to as the Krylov decomposition of $\mathcal{H}$ relative to A and g.

The invariance relations presented in Chapter 4 require some modification, as do their proofs. More specifically, the proof of Lemma 4.3.1 relies on topological aspects of continuity that cannot be used here as the operator is no longer guaranteed to be continuous. As such, further assumptions on the operator and Krylov space are necessary to ensure the invariance still hold at a suitably general level. To make these assumptions clear, the part of the operator A in the closed Krylov subspace is defined explicitly.

Definition 5.3.1. Let A be a closed, densely defined linear operator on Hilbert space $\mathcal{H}$, and let $g \in \mathcal{C}^\infty(A)$. Then the part of A on the closed Krylov subspace $\overline{\mathcal{K}(A,\, g)}$ is $\widetilde{A}$, i.e.,

$$\widetilde{A} := A|_{\overline{\mathcal{K}(A,\, g)} \cap \mathcal{D}(A)}\ . \tag{5.8}$$

Similarly, the part of the adjoint A^* on the closed Krylov subspace is $\widetilde{A^*}$, i.e.,

$$\widetilde{A^*} := A^*|_{\overline{\mathcal{K}(A,\, g)} \cap \mathcal{D}(A^*)}\ . \tag{5.9}$$

.

Remark 5.3.2. One may show that the operator $\widetilde{A}$ in Definition 5.3.1 is a *closed* operator (and so too is $\widetilde{A^*}$).

Indeed, take a graph norm convergent sequence $\Big((z_n, \widetilde{A}z_n)\Big)_{n\in\mathbb{N}} \subset G(\widetilde{A})$ (see Section B for details on graph spaces). Then

$$\begin{aligned} z_n &\xrightarrow{\|\cdot\|_{\mathcal{H}}} z \\ \widetilde{A}z_n &\xrightarrow{\|\cdot\|_{\mathcal{H}}} v\,. \end{aligned}$$

As $z_n \in \overline{\mathcal{K}\,(A,\, g)}$, it follows that $z \in \overline{\mathcal{K}\,(A,\, g)}$ too. As $\widetilde{A}$ is a restriction of A, one has that $\widetilde{A}u = Au$ for all $u \in \mathcal{D}(\widetilde{A})$. Therefore, $\widetilde{A}z_n = Az_n \xrightarrow{\|\cdot\|_{\mathcal{H}}} v$ and the closure of A guarantees that $z \in \mathcal{D}\,(A)$ and $v = Az$. So $z \in \overline{\mathcal{K}\,(A,\, g)} \cap \mathcal{D}\,(A) = \mathcal{D}(\widetilde{A})$ and therefore $Az = \widetilde{A}z = v$. Thus $\widetilde{A}$ is closed.

By a similar argument for $\widetilde{A^*}$, one proves that this operator is also closed.

Proposition 5.3.3. *Given a closed, densely defined operator $A : \mathcal{H} \to \mathcal{H}$ (with domain $\mathcal{D}\,(A)$) and vector $g \in \mathcal{C}^\infty\,(A)$, then the following invariance relation for the adjoint operator holds.*

$$A^* \left[\mathcal{K}\,(A,\, g)^\perp \cap \mathcal{D}\,(A^*)\right] \subset \mathcal{K}\,(A,\, g)^\perp\ . \tag{5.10}$$

Proof. For arbitrary $z \in \overline{\mathcal{K}\,(A,\, g)}$ take a sequence $(z_n)_n \subset \mathcal{K}\,(A,\, g)$, so that $z_n \xrightarrow{\|\cdot\|_{\mathcal{H}}} z \in \overline{\mathcal{K}\,(A,\, g)}$, and let $v \in \mathcal{K}\,(A,\, g)^\perp \cap \mathcal{D}\,(A^*)$, so that

$$0 = \langle Az_n,\, v\rangle = \langle z_n,\, A^*v\rangle\ .$$

Then $\langle z,\, A^*v\rangle = \lim_{n\to\infty} \langle z_n,\, A^*v\rangle = 0$. This implies that $A^*v \in \mathcal{K}\,(A,\, g)^\perp$. □

Lemma 5.3.4. *Let $A : \mathcal{H} \to \mathcal{H}$ be a closed, densely defined operator on Hilbert space $\mathcal{H}$ (with domain $\mathcal{D}\,(A)$), with a vector $g \in \mathcal{C}^\infty\,(A)$. If $\mathcal{K}\,(A,\, g)$ is a core of $\widetilde{A}$, then the following generalised invariance relation holds.*

$$A\left[\overline{\mathcal{K}\,(A,\, g)} \cap \mathcal{D}\,(A)\right] \subset \overline{\mathcal{K}\,(A,\, g)}\,. \tag{5.11}$$

Proof. Let $z \in \overline{\mathcal{K}(A,\, g)} \cap \mathcal{D}(A)$. Hence $(z, Az) \in G(A)$ but also $(z, Az) = (z, \widetilde{A}z) \in G(\widetilde{A})$, where $\widetilde{A}$ is defined by (5.8) and is closed by Remark 5.3.2. By the assumption that $\mathcal{K}(A,\, g)$ is a core of $\widetilde{A}$, the operator $A' := \widetilde{A}|_{\mathcal{K}(A,\, g)}$ satisfies

$$\overline{G(A')} = G(\widetilde{A})\,.$$

As such, one may deduce that there is a sequence of approximants $(z_n, Az_n) \in G(A')$ of (z, Az) meaning that

$$\begin{aligned} \mathcal{K}(A,\, g) \ni z_n &\xrightarrow{\|\cdot\|_{\mathcal{H}}} z \\ Az_n &\xrightarrow{\|\cdot\|_{\mathcal{H}}} Az\,. \end{aligned}$$

This implies that $Az \in \overline{A\mathcal{K}(A,\, g)} \subset \overline{\mathcal{K}(A,\, g)}$. □

Remark 5.3.5. As stated in the proof of Lemma 5.3.4, the assumption that $\mathcal{K}(A,\, g)$ is a core of $\widetilde{A}$ is *equivalent* to having $\overline{G(A')} = G(\widetilde{A})$ (Definition 5.2.6). To have that $\overline{G(A')} = G(\widetilde{A})$ it is *necessary* that $\mathcal{K}(A,\, g)$ is dense in $\mathcal{D}(\widetilde{A}) = \overline{\mathcal{K}(A,\, g)} \cap \mathcal{D}(A)$, which is certainly satisfied as $\mathcal{K}(A,\, g) \subset \mathcal{D}(A)$. However, the density alone of $\mathcal{K}(A,\, g)$ in $\overline{\mathcal{K}(A,\, g)} \cap \mathcal{D}(A)$ is *not* sufficient in general to guarantee that it is a core, unless $A \in \mathscr{B}(\mathcal{H})$. Therefore the density of $\mathcal{K}(A,\, g)$ in $\mathcal{D}(\widetilde{A})$ is *not* tantamount to $\mathcal{K}(A,\, g)$ being a core of $\widetilde{A}$, so this assumption in Lemma 5.3.4 is required in its proof.

The concept of Krylov reducibility is also generalised in what follows.

Definition 5.3.6. Given a closed, densely defined linear operator A in Hilbert space $\mathcal{H}$ and $g \in \mathcal{C}^\infty(A)$, one says that the operator A is *$\mathcal{K}(A,\, g)$-reduced* when $\overline{\mathcal{K}(A,\, g)}$ and $\mathcal{K}(A,\, g)^\perp$ are invariant under A, i.e., $A(\overline{\mathcal{K}(A,\, g)} \cap \mathcal{D}(A)) \subset \overline{\mathcal{K}(A,\, g)}$ and $A(\mathcal{K}(A,\, g)^\perp \cap \mathcal{D}(A)) \subset \mathcal{K}(A,\, g)^\perp$. This is referred to as *(generalised) $\mathcal{K}(A,\, g)$-reducibility* of A, or simply *(generalised) Krylov reducibility* where no confusion arises.

Remark 5.3.7. The proofs of Lemma 4.3.4 and Proposition 4.3.7 are not suitable for extension to the entire class of closed, densely defined operators.

Concerning Lemma 4.3.4, the lack of a similar proof stems from the fact that given a closed subspace $\mathcal{V} \subset \mathcal{H}$, one may *not* always say that $\mathcal{V} \cap \mathcal{D}(A)$

is dense in $\mathcal{V}$ (e.g., think of $\mathcal{V} = \operatorname{span}\{v\}$ for some $v \in \mathcal{H} \setminus \mathcal{D}(A)$). And yet a similar argument of the equivalence between the following

(i) $A(\mathcal{V} \cap \mathcal{D}(A)) \subset \mathcal{V}$, $A(\mathcal{V}^\perp \cap \mathcal{D}(A)) \subset \mathcal{V}^\perp$, and

(ii) $A^*(\mathcal{V} \cap \mathcal{D}(A^*)) \subset \mathcal{V}$, $A^*(\mathcal{V}^\perp \cap \mathcal{D}(A^*)) \subset \mathcal{V}^\perp$

would require density of $\mathcal{V}\cap\mathcal{D}(A)$, $\mathcal{V}\cap\mathcal{D}(A^*)$ *in* $\mathcal{V}$, *and* $\mathcal{V}^\perp\cap\mathcal{D}(A)$, $\mathcal{V}^\perp\cap\mathcal{D}(A^*)$ *in* $\mathcal{V}^\perp$.

The proof to Proposition 4.3.7 relies on an equivalence between (i) and (ii) above, and so is unsuitable for modification to the entire class of closed, densely defined operators.

Finally the concept of the Krylov intersection is suitably generalised. In the following section, again it is shown that this is the intrinsic operator-theoretic mechanism that guarantees Krylov solvability of the linear inverse problem.

Definition 5.3.8. Given a closed, densely defined linear operator A on Hilbert space $\mathcal{H}$ and a vector $g \in \mathcal{C}^\infty(A)$, the intersection

$$\overline{\mathcal{K}(A,\,g)} \cap \Big[A(\mathcal{K}(A,\,g)^\perp \cap \mathcal{D}(A))\Big] \tag{5.12}$$

is called the *(generalised) Krylov intersection*, and is denoted by $\widetilde{\mathscr{I}_{\mathcal{K}}}(A,\,g)$.

For a given closed, densely defined operator $A : \mathcal{H} \to \mathcal{H}$ and $g \in \mathcal{C}^\infty(A)$, the consequence of A being Krylov reducible guarantees that $\widetilde{\mathscr{I}_{\mathcal{K}}}(A,\,g) = \{0\}$.

5.4 Krylov Solvability

In this Section, the theorems and lemmas regarding Krylov solvability, as introduced in Chapter 4, Section 4.4, are suitably generalised to the closed, densely defined operator class. Some additional assumptions, made explicit in the modified statements, are needed owing to the possible unboundedness of the operator class.

To begin with, the appropriate analogue of Proposition 4.4.2 is presented. Density of $A\mathcal{K}(A, g)$ in $\overline{\mathcal{K}(A, g)}$ is still a *necessary* condition for Krylov solvability, under assumptions on the core of $\widetilde{A}$, that become *necessary and sufficient* if $A^{-1} \in \mathscr{B}(\mathcal{H})$.

Proposition 5.4.1. *Let $A : \mathcal{H} \to \mathcal{H}$ be a closed, densely defined, injective linear operator on Hilbert space $\mathcal{H}$, and let $f \in \mathcal{D}(A)$ be the solution to $Af = g$, given $g \in \operatorname{ran} A \cap C^\infty(A)$. One has the following.*

(i) If $f \in \overline{\mathcal{K}(A, g)}$ and $\mathcal{K}(A, g)$ is a core of $\widetilde{A}$, then $A\mathcal{K}(A, g)$ is dense in $\overline{\mathcal{K}(A, g)}$.

(ii) If A is invertible with an everywhere defined bounded inverse and $A\mathcal{K}(A, g)$ is dense in $\overline{\mathcal{K}(A, g)}$, then $f \in \overline{\mathcal{K}(A, g)}$.

Proof. Starting with (i), by assumption $f \in \overline{\mathcal{K}(A, g)}$. $A\mathcal{K}(A, g) \subset \mathcal{K}(A, g)$ implies $\overline{A\mathcal{K}(A, g)} \subset \overline{\mathcal{K}(A, g)}$. As A is a closed operator and $\mathcal{K}(A, g)$ is a core for $\widetilde{A}$, there exists $(f_n)_{n\in\mathbb{N}}$ in $\mathcal{K}(A, g)$ such that $f_n \xrightarrow{\|\cdot\|_{\mathcal{H}}} f$ and $\mathcal{K}(A, f) \ni Af_n \xrightarrow{\|\cdot\|_{\mathcal{H}}} Af = g$, i.e., $f_n \xrightarrow{\|\cdot\|_{G(A)}} f$. Then $g \in \overline{A\mathcal{K}(A, g)}$, and

$$\operatorname{span}\left\{A^k g \,|\, k \in \mathbb{N}_0\right\} \subset \overline{A\mathcal{K}(A, g)},$$

so that $\overline{\mathcal{K}(A, g)} \subset \overline{A\mathcal{K}(A, g)}$. So $\overline{A\mathcal{K}(A, g)} = \overline{\mathcal{K}(A, g)}$.

For (ii) assume that A has bounded, everywhere defined inverse, and that $A\mathcal{K}(A, g)$ is dense in $\overline{\mathcal{K}(A, g)}$. There is a sequence of approximants in $A\mathcal{K}(A, g)$ to the vector $g \in \overline{\mathcal{K}(A, g)}$, namely $(Av_n)_{n\in\mathbb{N}}$ such that $v_n \in \mathcal{K}(A, g) \subset \mathcal{D}(A)$ for all $n \in \mathbb{N}$. If $A^{-1} \in \mathscr{B}(\mathcal{H})$ then $\|Av_n - g\|_{\mathcal{H}} \to 0$ implies $\|v_n - f\|_{\mathcal{H}} \to 0$, as $\|v_n - f\|_{\mathcal{H}} = \|A^{-1}(Av_n - g)\|_{\mathcal{H}} \leq \|A^{-1}\|_{\text{op}} \|Av_n - g\|_{\mathcal{H}} \to 0$. So v_n converges to $f \in \overline{\mathcal{K}(A, g)}$. □

A *sufficient* condition to ensure Krylov solvability of the well-defined linear inverse problem is the Krylov reducibility, coupled with the requirement that the orthogonal projection of the solution vector f onto the closed Krylov subspace remains within the domain of the operator A. This latter requirement is necessary in the proof of the following proposition to ensure that all the operations involving A are properly defined. Therefore this additional restriction, much like the condition of $\mathcal{K}(A,\,g)$ being a core for $\widetilde{A}$ to guarantee invariance of $\overline{\mathcal{K}(A,\,g)}$, is inescapable owing to the possible unboundedness of the operator.

Proposition 5.4.2. *Let A be a closed, densely defined, injective linear operator on Hilbert space $\mathcal{H}$, and let $f \in \mathcal{D}(A)$ be the solution to $Af = g$, given $g \in \mathrm{ran}A \cap \mathcal{C}^\infty(A)$. Let the linear operator $P_\mathcal{K} : \mathcal{H} \to \mathcal{H}$ be the orthogonal projection operator onto the space $\overline{\mathcal{K}(A,\,g)}$. If A is $\mathcal{K}(A,\,g)$-reduced and $P_\mathcal{K} f \in \mathcal{D}(A)$, then $f \in \overline{\mathcal{K}(A,\,g)}$. In addition, if A is self-adjoint, and $\mathcal{K}(A,\,g)$ is a core of $\widetilde{A}$, then $Af = g$ implies that $f \in \overline{\mathcal{K}(A,\,g)}$.*

Proof. Decompose the solution $f = P_\mathcal{K} f + (\mathbb{1} - P_\mathcal{K})f$. As $f \in \mathcal{D}(A)$ and by assumption $P_\mathcal{K} f \in \mathcal{D}(A)$, this implies that $(\mathbb{1} - P_\mathcal{K}) \in \mathcal{D}(A)$. So

$$Af = g = AP_\mathcal{K} f + A(\mathbb{1} - P_\mathcal{K})f\,.$$

From the definition of Krylov reducibility $AP_\mathcal{K} f \in \overline{\mathcal{K}(A,\,g)}$, so that $A(\mathbb{1} - P_\mathcal{K})f \in \overline{\mathcal{K}(A,\,g)}$. Again using $\mathcal{K}(A,\,g)$-reducibility of A, $A(\mathbb{1} - P_\mathcal{K})f \in \mathcal{K}(A,\,g)^\perp$, so $(\mathbb{1} - P_\mathcal{K})f = 0$ owing to injectivity. Then $f \in \overline{\mathcal{K}(A,\,g)}$.

By Proposition 5.3.3 and Lemma 5.3.4, if additionally A is self-adjoint and $\mathcal{K}(A,\,g)$ is a core of $\widetilde{A}$, then A is $\mathcal{K}(A,\,g)$-reduced. □

Remark 5.4.3. At this stage, a concrete example of a self-adjoint operator such that $\mathcal{K}(A,\,g)$ is *not* a core for the domain restricted operator $\widetilde{A}$ would be interesting to have. This would provide a beautiful contrast to the bounded scenario ($A \in \mathscr{B}(\mathcal{H})$) as there, $\mathcal{K}(A,\,g)$ is *always* a core for $\widetilde{A}$.

Unfortunately a counterpart to Proposition 4.4.4 for self-adjoint operators using spectral integrals does not work in this situation. The proof may not be suitably modified as the Stone-Weierstrass theorem on locally compact spaces requires the construction of an algebra of functions that *vanish* at infinity. Clearly, this is *not* satisfied by the class of polynomial functions on $\mathbb{R}$. The construction may still be made when considering rational Krylov spaces, and this will be seen later.

The fundamental relationship between the triviality of the generalised Krylov intersection and the Krylov solvability of the linear inverse problem still holds, under certain restrictions. Certainly, the following proposition shows that the triviality of the Krylov intersection is still the intrinsic mechanism capturing Krylov solvability.

Proposition 5.4.4. *Let $A : \mathcal{H} \to \mathcal{H}$ be a closed, densely defined, injective operator on a Hilbert space $\mathcal{H}$, and let $f \in \mathcal{D}(A)$ be the solution to $Af = g$, given $g \in \operatorname{ran}A \cap \mathcal{C}^\infty(A)$. Let the linear operator $P_\mathcal{K} : \mathcal{H} \to \mathcal{H}$ be the orthogonal projection operator onto the space $\overline{\mathcal{K}(A, g)}$.*

(i) If $P_\mathcal{K} f \in \mathcal{D}(A)$ and $\widetilde{\mathscr{I}_\mathcal{K}}(A, g) = \{0\}$, then $f \in \overline{\mathcal{K}(A, g)}$.

(ii) If A has an everywhere defined, bounded inverse on $\mathcal{H}$, $\mathcal{K}(A, g)$ is a core of $\widetilde{A}$, and $f \in \overline{\mathcal{K}(A, g)}$; then $\widetilde{\mathscr{I}_\mathcal{K}}(A, g) = \{0\}$.

Proof. Working on part (i), by assumption $f \in \mathcal{D}(A)$ and $P_\mathcal{K} f \in \mathcal{D}(A)$, so that $(\mathbb{1} - P_\mathcal{K})f \in \mathcal{D}(A)$. As in the proof of Proposition 5.4.2, one has $A(\mathbb{1} - P_\mathcal{K})f \in \overline{\mathcal{K}(A, g)}$ because $AP_\mathcal{K} f \in \overline{\mathcal{K}(A, g)}$. This, together with $\widetilde{\mathscr{I}_\mathcal{K}}(A, g) = \{0\}$ and $A(\mathbb{1} - P_\mathcal{K})f \in A\left[\mathcal{K}(A, g)^\perp \cap \mathcal{D}(A)\right]$, ensures that $A(\mathbb{1} - P_\mathcal{K})f = 0$. From the injectivity, $(\mathbb{1} - P_\mathcal{K})f = 0$ and $f \in \overline{\mathcal{K}(A, g)}$.

Considering part (ii), one has $f = A^{-1}g$, and $f \in \overline{\mathcal{K}(A, g)} \cap \mathcal{D}(A)$. Then take some $z \in \widetilde{\mathscr{I}_\mathcal{K}}(A, g)$ and let $z = Aw$ for some $w \in \mathcal{K}(A, g)^\perp \cap \mathcal{D}(A)$. From Proposition 5.4.1 (i), $A\mathcal{K}(A, g)$ is dense in $\overline{\mathcal{K}(A, g)}$. So there is some sequence $(v_n)_{n\in\mathbb{N}} \subset \mathcal{K}(A, g) \subset \mathcal{D}(A)$ such that $Av_n \to z = Aw$ in the $\mathcal{H}$-norm. $A^{-1} \in \mathscr{B}(\mathcal{H})$ implies $\|v_n - w\|_\mathcal{H} = \|A^{-1}A(v_n - w)\|_\mathcal{H} \leq \|A^{-1}\|_{\text{op}} \|A(v_n - w)\|_\mathcal{H} \to 0$. Therefore, $v_n \xrightarrow{\|\cdot\|_\mathcal{H}} w$ and

$$w \in \mathcal{K}(A, g)^\perp \cap \mathcal{D}(A) \ , \quad v_n \in \mathcal{K}(A, g) \ .$$

From the above, v_n and w are in othogonally complementary spaces, so that

$$0 = \lim_{n\to\infty} \|v_n - w\|^2_\mathcal{H} = \lim_{n\to\infty} \left(\|v_n\|^2_\mathcal{H} + \|w\|^2_\mathcal{H}\right) = 2\,\|w\|^2_\mathcal{H}$$

so $w = 0$ which implies $z = 0$. □

Remark 5.4.5. For self-adjoint operators on $\mathcal{H}$, the assumption of a core in Proposition 5.4.2 is no longer required for the Krylov solvability of the linear inverse problem. Actually, the triviality of the Krylov intersection, along with $P_{\mathcal{K}} f \in \mathcal{D}(A)$, is all that is needed for Krylov solvability.

5.4.1 Krylov solutions in the lack of injectivity

Counterparts of Krylov solvability of the linear inverse problem in the lack of injectivity are presented with little or no modifications to the underlying proofs from Chapter 4.

Proposition 5.4.6. *Let A be a closed, densely defined, linear operator on Hilbert space $\mathcal{H}$, and let $Af = g$ be the associated linear inverse problem with $g \in \operatorname{ran} A \cap \mathcal{C}^\infty(A)$. If $\ker A \subset \ker A^*$, then there exists at most one solution $f \in \overline{\mathcal{K}(A, g)}$. In particular this statement holds true if A is normal.*

Proof. Identical to Proposition 4.4.13. □

Proposition 5.4.7. *Let A be a closed, densely defined linear operator on Hilbert space $\mathcal{H}$ and let $g \in \operatorname{ran} A \cap \mathcal{C}^\infty(A)$. Let $P_{\mathcal{K}} : \mathcal{H} \to \mathcal{H}$ be the orthogonal projection operator onto $\overline{\mathcal{K}(A, g)}$. If A is $\mathcal{K}(A, g)$-reduced, and if $f_\circ \in \mathcal{D}(A)$ satisfies $Af_\circ = g$ is such that $P_{\mathcal{K}} f_\circ \in \mathcal{D}(A)$; then $f := P_{\mathcal{K}} f_\circ$ is a Krylov solution.*

Proof. Owing to the same argument in the proof of Proposition 5.4.2, one has $A(\mathbb{1} - P_{\mathcal{K}})f_\circ = 0$. Then $AP_{\mathcal{K}} f_\circ = g$, i.e. $f := P_{\mathcal{K}} f_\circ \in \mathcal{D}(A)$ is a Krylov solution. □

The previous two propositions combine to give the following corollary concerning the uniqueness of Krylov solutions for self-adjoint linear inverse problems.

Corollary 5.4.8. *Let A be a self-adjoint linear operator on the Hilbert space $\mathcal{H}$. If the linear inverse problem $Af = g$, with $g \in \operatorname{ran} A \cap \mathcal{C}^\infty(A)$, has a solution $f \in \overline{\mathcal{K}(A, g)}$; then f is the unique Krylov solution.*

Proof. As $f \in \overline{\mathcal{K}(A, g)}$ is a solution to $Af = g$, immediately from Proposition 5.4.6 one has that f is unique in $\overline{\mathcal{K}(A, g)}$. □

5.4.2 Some remarks on rational Krylov solvability for self-adjoint operators

In this Section some preliminary results regarding *rational* Krylov solvability are developed for the specific class of self-adjoint operators. The theorems developed here show that, in the case of unbounded operators, from an approximation standpoint it may be advantageous to consider general *rational* approximations rather than standard polynomials. Additionally, the restriction that $g \in \mathcal{C}^{\infty}(A)$ may be dropped. The practical drawback of such an approach is that there is an extra computational cost in calculating the resolvent function $\mathcal{R}(A, \xi)$ for some $\xi \in \rho(A)$.

In this discussion, a solution f to $Af = g$ belonging to the closure of a rational Krylov subspace associated with A, g and a sequence Ξ, i.e. $f \in \overline{\mathcal{K}^{\Xi}(A, g)}$, is referred to as a *rational Krylov solution.* Informally, one refers to the linear inverse problem being *rationally Krylov solvable* should there exist at least one solution $f \in \overline{\mathcal{K}^{\Xi}(A, g)}$.

Of particular interest here is Corollary 5.4.11 that is a result of relevance for the so-called 'shift and invert method' developed in [58, 94].

Theorem 5.4.9. *Let A be a self-adjoint operator, with spectral measure $\mathbf{E}(t)$, and scalar measure $\mu_g(t) := \langle g, \mathbf{E}(t)\, g\rangle$ for given $g \in \mathcal{H}$. Let $(\xi_n)_{n\in\mathbb{N}} \subset \mathbb{C}$ be a sequence such that*

(i) ξ_n is in the resolvent of A, for all $n \in \mathbb{N}$,

(ii) the set $\{\xi_n\}_{n\in\mathbb{N}}$ is closed under complex conjugation.

Consider $\mathcal{I}$, the collection of all finite index sets generated from $\mathbb{N}$. Then the subset $B \subset C_0(\sigma(A), \mathbb{C})$ defined by

$$B = \operatorname{span}\left\{ \prod_{n\in I} \frac{1}{z - \xi_n}\,;\ I \in \mathcal{I} \right\} \tag{5.13}$$

is dense in $L^2(\sigma(A), \mu_g)$.

Proof. As $\mu_g(\mathbb{R}) = \|g\|^2_{\mathcal{H}} < \infty$, obviously μ_g is a regular Borel measure.

B is an involutive subalgebra of $C_0(\sigma(A), \mathbb{C})$: as it contains both $z \mapsto (z - \xi_n)^{-1}$ and $z \mapsto (z - \overline{\xi_n} =^{-1}$, and is clearly closed under sums and products.

Furthermore, B separates points in $\sigma(A)$ and B vanishes nowhere on $\sigma(A)$ in the sense of Definitions C.2.1 and C.2.2. $\sigma(A)$ is closed in $\mathbb{C}$ and therefore locally compact. So, from the Stone-Weierstrass theorem for locally compact spaces (Theorem C.2.4), one has that $\overline{B}^{\|\cdot\|_\infty} = C_0(\sigma(A), \mathbb{C})$. Moreover, from $\overline{C_c(\sigma(A), \mathbb{C})}^{\|\cdot\|_2} = L^2(\sigma(A), \mu_g)$ [79, Theorem 3.14], and $C_c(\sigma(A), \mathbb{C}) \subset C_0(\sigma(A), \mathbb{C})$, one has also $\overline{C_0(\sigma(A), \mathbb{C})}^{\|\cdot\|_2} = L^2(\sigma(A), \mu_g)$.

Given $u \in L^2(\sigma(A), \mu_g)$ and $\varepsilon > 0$ there exists $h \in C_0(\sigma(A), \mathbb{C})$ such that

$$\|(h(A) - u(A))g\|_{\mathcal{H}}^2 = \int_{\sigma(A)} |h(t) - u(t)|^2 \, \mathrm{d}\mu_g(t) < \frac{\varepsilon^2}{2},$$

and there exists some $p \in B$ such that

$$\begin{aligned} \|(p(A) - h(A))g\|_{\mathcal{H}}^2 &= \int_{\sigma(A)} |p(t) - h(t)|^2 \, \mathrm{d}\mu_g(t) \\ &\leq \|p(t) - h(t)\|_{L^\infty(\sigma(A), \mu_g)}^2 \|g\|_{\mathcal{H}}^2 < \frac{\varepsilon^2}{2}. \end{aligned}$$

Therefore,

$$\|(p(A) - u(A))g\|_{\mathcal{H}}^2 < \varepsilon^2,$$

which implies $\overline{B}^{\|\cdot\|_2} = L^2(\sigma(A), \mu_g)$. □

Corollary 5.4.10. *Let $A : \mathcal{H} \to \mathcal{H}$ be a self-adjoint, injective operator, and let $(\xi_n)_{n \in \mathbb{N}}$ and B as in Theorem 5.4.9. Then if $Af = g$ with $g \in \mathrm{ran} A$ one has*

$$f \in \overline{\operatorname{span}\left\{ \prod_{n \in I} (A - \xi_n \mathbb{1})^{-1} g;\ I \in \mathcal{I} \right\}}.$$

Proof. $B = \operatorname{span}\left\{\prod_{n \in S}(z - \xi_n)^{-1} g\right\}_{S \in \mathcal{S}}$ is dense in $L^2(\sigma(A), \mu_g)$ from Theorem 5.4.9, and clearly $\frac{1}{t} \in L^2(\sigma(A), \mu_g)$, as $\|A^{-1}g\|_{\mathcal{H}}^2 = \|f\|_{\mathcal{H}}^2 < \infty$. □

Corollary 5.4.11. *Let $A : \mathcal{H} \to \mathcal{H}$ be a self-adjoint, injective operator such that there exists $\xi \in \rho(A) \cap \mathbb{R}$, and let $\Xi \equiv (\xi_n)_{n \in \mathbb{N}}$ with $\xi_n = \xi$ for all $n \in \mathbb{N}$. Then the solution f to $Af = g$ with $g \in \mathrm{ran} A$ belongs to the space $\overline{\mathcal{K}^{\Xi}(A,\, g)}$,* *i.e.,*

$$f \in \overline{\operatorname{span}\left\{\prod_{n=1}^{m}(A-\xi\mathbb{1})^{-1}g\,;\,m\in\mathbb{N}\right\}\oplus\operatorname{span}\{g\}}\,.$$

Proof. $B = \operatorname{span}\{\prod_{n=1}^{m}(z-\xi_n)^{-1}\,|\,m\in\mathbb{N}\}$ satisfies the conditions of Theorem 5.4.9, so B is dense in $L^2(\sigma(A),\mu_g)$ and the approximation result follows. □

Remark 5.4.12. The rational Krylov space $\mathcal{K}^{\Xi}(A,\,g)$ from Corollary 5.4.11 will successfully approximate *any* function of the operator applied to g, i.e., $h(A)g$, provided that $h\in L^2(\sigma(A),\mu_g)$. Indeed $\|h(A)g-p(A)g\|^2_{\mathcal{H}} = \|h(t)-(t)\|_{L^2(\sigma(A),\mu_g)}$ for $p\in B$. This is also *regardless* of whether A is injective or not, as injectivity is only used to ensure that $t^{-1}\in L^2(\sigma(A),\mu_g)$.

The same comment may be made about the space considered in Corollary 5.4.10: again the injectivity requirement here is only needed to ensure $t^{-1}\in L^2(\sigma(A),\mu_g)$.

Remark 5.4.13. The $\mathcal{H}$-norm closures of the subspaces generated in Theorem 5.4.9, Corollary 5.4.10 and Corollary 5.4.11 are identical to the closure of the *polynomial* Krylov subspace for the class of *bounded* operators $A\in\mathcal{B}(\mathcal{H})$. This is an immediate result from the isomorphisms $\overline{\mathcal{K}(A,\,g)}\cong L^2(\sigma(A),\mu_g)$ and $\overline{\mathcal{K}^{\Xi}(A,\,g)}\cong L^2(\sigma(A),\mu_g)$, where the former was already obtained in the discussion right after Proposition 4.4.4, and the latter follows from a completely analogous reasoning from the proof of Corollary 5.4.10.

Chapter 6

Conjugate-gradients for Unbounded Operators

6.1 Introduction

A specific application of the solution(s) to linear inverse problems using the class of Krylov subspace methods known as conjugate-gradient methods, is discussed in the setting of a general *self-adjoint, positive operator*. The analysis contained herein has been inspired by the deep work of Nemirovskiy and Polyak [64] where they considered the convergence and its rate for conjugate-gradient style methods in the framework of *bounded* operators.

In this Chapter the analysis is generalised further to the setting where A may be unbounded. Specifically, the general setting considered here is for the solvable linear inverse problem on Hilbert space $\mathcal{H}$

$$Af = g\,, \quad g \in \mathrm{ran}A\,, \tag{6.1}$$

where $A : \mathcal{H} \to \mathcal{H}$ is a self-adjoint (therefore closed), with the positivity condition, i.e., $\langle \psi, A\psi \rangle \geq 0$, equivalently $A \geq \mathbb{O}$, for all $\psi \in \mathcal{D}(A)$.

At an informal level, the algorithm of conjugate-gradients is a minimisation scheme in a chosen semi-norm defined by the operator A and a parameter $\theta \geq 0$. Analogous to the presentation in Chapter 2, the scheme works by taking a suitable initial guess $f^{[0]}$ and datum g, and picks the N-th approximation according to the prescription

$$f^{[N]} := \underset{h\in\{f^{[0]}\}+\mathcal{K}_N(A,\mathfrak{R}_0)}{\operatorname{argmin}} \left\|A^{\theta/2}(h-\mathcal{P}_{\mathcal{S}}f^{[0]})\right\|_{\mathcal{H}} . \tag{6.2}$$

Here, $\mathcal{P}_{\mathcal{S}}f^{[0]}$ is the closest to $f^{[0]}$ solution to $Af = g$, and $\mathfrak{R}_0 = Af^{[0]} - g$. The procedure is described formally in Definition 6.2.7. The main result of this Chapter is that the *convergence of the iterates* $f^{[N]} \to \mathcal{P}_{\mathcal{S}}f^{[0]}$ *is controllable in a suitable sense.* The exact sense is made clear in the formulation of the main result, Theorem 6.4.1, and the remarks that follow.

This is indeed novel, for the case of A being an unbounded operator has only recently been considered from special perspectives. The view of the existence of Krylov solvability was taken in Olver [69] where the author considered the GMRES scheme applied to the linear inverse problem from a first order derivative operator. For the conjugate-gradient scheme, Gilles and Townsend [35] considered the case in which the operator A was a second order differential operator, suitably regularised to ensure that it had an everywhere defined bounded inverse. In light of these studies, the convergence theory from the most *general*, abstract perspective is missing.

Therefore it is possible that one simultaneously has an unbounded linear operator, with $0 \in \sigma(A)$. Moreover, 0 may be an accumulation point of the spectrum, making the linear inverse problem ill-posed (as opposed to having the properties of well-posedness).

Remark 6.1.1. In the context of self-adjoint operators (not necessarily bounded), in infinite-dimensional Hilbert space, the following properties are all equivalent.

(i) $0 \in \sigma(A)$ and 0 is *not* an isolated point in $\sigma(A)$,

(ii) The range is *not* closed in $\mathcal{H}$,

(iii) The operator has unbounded inverse on the range.

These occurrences are possible only if $\dim \mathcal{H} = \infty$.

6.2 Definitions, set-up, and comments

To begin with, some definitions and notations are made before the general algorithm is described. Many accompanying remarks are made along the way, owing to the subtleties in dealing with the unbounded operator scenario. Here, and in what follows, $\mathbf{E}$ will denote the projection valued measure associated with the self-adjoint operator A (see Appendix B). The quantity $\mathrm{d}\left\langle x, \mathbf{E}\left(t\right)x\right\rangle$ denotes the corresponding scalar measure associated to a vector $x \in \mathcal{H}$. These measures are supported on the spectrum of A.

For the sake of convenience, the following sets of polynomials are defined for use in this Chapter.

Definition 6.2.1. Let $t \in [0, \infty)$, then one may define the following polynomial collections on the positive real line.

$$\mathbb{P}([0,\infty)) := \{\text{polynomials } p(t),\, t \in [0,\infty)\} \tag{6.3}$$

$$\mathbb{P}_N := \{p \in \mathbb{P}([0,\infty)) \,|\, \deg p \leq N\} \tag{6.4}$$

$$\mathbb{P}_N^{(1)} := \{p \in \mathbb{P}_N \,|\, p(0) = 1\}\,. \tag{6.5}$$

Following this, the solution manifold is defined as follows.

Definition 6.2.2. Consider the solvable linear inverse problem (6.1). Then the solution manifold $\mathcal{S}$ is the set of points

$$\mathcal{S} := \{f \in \mathcal{D}\left(A\right) \mid Af = g\}\,, \tag{6.6}$$

and the operator $\mathcal{P}_{\mathcal{S}} : \mathcal{H} \to \mathcal{S}$ is the projection map onto the manifold $\mathcal{S}$.

Remark 6.2.3. As the operator A is closed, the kernel too is closed, and therefore the solution manifold is a closed convex set. It is therefore known that the operator $\mathcal{P}_{\mathcal{S}}$ is well-defined and produces, for any $v \in \mathcal{H}$, the closest to v point in $\mathcal{S}$ [10, Chapter 5]. Moreover, $\mathcal{S} \neq \emptyset$ as $g \in \mathrm{ran} A$.

The following two lemmas are technical facts to facilitate the further discussion of various definitions and remarks of this Section. They may be skipped as to not interrupt the flow of the current section, without causing any significant confusion.

Lemma 6.2.4. *Let $A : \mathcal{H} \to \mathcal{H}$ be a self-adjoint, positive operator in $\mathcal{H}$. Then*

(i) $y \in \ker A$ if and only if $\|\mathbf{E}\left((0,\infty)\right) y\|_{\mathcal{H}} = 0$,

(ii) and $y \in (\ker A)^{\perp}$ if and only if $\|\mathbf{E}\left(\{0\}\right) y\|_{\mathcal{H}} = 0$.

Proof. Starting with part (i), assuming that $\|\mathbf{E}\left((0,\infty)\right) y\|_{\mathcal{H}} = 0$ and letting $\varepsilon > 0$, one immediately sees that

$$\begin{aligned}
\|Ay\|_{\mathcal{H}}^2 &= \int_{[0,\varepsilon)} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle + \int_{[\varepsilon,\infty)} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle \\
&\leq \varepsilon^2 \left\langle y,\, \mathbf{E}\left([0,\varepsilon)\right) y\right\rangle + \int_{[\varepsilon,\infty)} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle \\
&= \varepsilon^2 \left\langle y,\, \mathbf{E}\left([0,\varepsilon)\right) y\right\rangle + \lim_{n\to\infty} \int_{[\varepsilon,n)} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle \\
&\leq \varepsilon^2 \left\langle y,\, \mathbf{E}\left([0,\varepsilon)\right) y\right\rangle + \lim_{n\to\infty} n^2 \left\langle y,\, \mathbf{E}\left((\varepsilon,n)\right) y\right\rangle \\
&= \varepsilon^2 \left\langle y,\, \mathbf{E}\left([0,\varepsilon)\right) y\right\rangle \leq \varepsilon^2 \|y\|_{\mathcal{H}}^2 \xrightarrow{\varepsilon\to 0} 0\,,
\end{aligned}$$

owing to the monotonicity property of the measure $\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle$ ensuring that $\left\langle y,\, \mathbf{E}\left((\varepsilon,n)\right) y\right\rangle \leq \left\langle y,\, \mathbf{E}\left((0,\infty)\right) y\right\rangle = 0$ for all $n \in \mathbb{N}$. The step $\int_{[\varepsilon,\infty)} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle = \lim_{n\to\infty} \int_{[\varepsilon,\infty)} t^2 \chi_{[\varepsilon,n)} \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle$ is justified by the Lebesgue monotone convergence theorem ([79, Theorem 1.26]). This proves the backward implication in (i). For the forward implication, let $y \in \ker A$, so that $\|Ay\|_{\mathcal{H}} = 0$. If, by contradiction, $\|\mathbf{E}\left((0,\infty)\right) y\|_{\mathcal{H}} > 0$, then

$$\begin{aligned}
0 &= \int_{\{0\}} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle + \int_{(0,\infty)} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle \\
&= \int_{(0,\infty)} t^2 \,\mathrm{d}\left\langle y,\, \mathbf{E}\left(t\right) y\right\rangle \,.
\end{aligned}$$

Yet, because $t^2 > 0$ and is continuous on $(0,\infty)$, the last integral cannot be zero. Thus, one must necessarily have $\|\mathbf{E}\left((0,\infty)\right) y\|_{\mathcal{H}} = 0$. This completes the proof of part (i).

For part (ii), consider the forward implication. By assumption $y \in (\ker A)^{\perp}$. Decompose y into $y = \mathbf{E}(\{0\})\, y + \mathbf{E}((0,\infty))\, y$. Clearly, $\mathbf{E}(\{0\})\, y \perp \mathbf{E}((0,\infty))\, y$. From part (i), $\mathbf{E}(\{0\})\, y \in \ker A$ (because $\mathbf{E}((0,\infty))\, \mathbf{E}(\{0\})\, y = \mathbf{E}(\emptyset)\, y = 0$). If, by contradiction, $\mathbf{E}(\{0\})\, y \neq 0$, then $y \notin (\ker A)^{\perp}$. For the backward implication, let $\|\mathbf{E}(\{0\})\, y\|_{\mathcal{H}} = 0$, so that by the same decomposition, one immediately obtains $y = \mathbf{E}((0,\infty))\, y$. Then, if $x \in \ker A$, by part (i) $x = \mathbf{E}(\{0\})\, x$, so that

$$\begin{aligned}\langle x,\, y\rangle &= \langle \mathbf{E}(\{0\})\, x,\, \mathbf{E}((0,\infty))\, y\rangle \\ &= \langle x,\, \mathbf{E}(\{0\})\, \mathbf{E}((0,\infty))\, y\rangle \\ &= \langle x,\, \mathbf{E}(\emptyset)\, y\rangle = 0\,.\end{aligned}$$

Therefore, $y \in (\ker A)^{\perp}$, and the proof of part (ii) is complete. □

Lemma 6.2.5. *Let $A : \mathcal{H} \to \mathcal{H}$ be a self-adjoint, positive operator in $\mathcal{H}$. Then the following hold true.*

(i) $\ker(A^{\theta}) = \ker A$ *for any* $\theta > 0$.

(ii) $\ker A$ *and* $(\ker A)^{\perp}$ *remain invariant under the operator* A^{θ} *for any* $\theta \geq 0$.

(iii) If a vector $v \in \mathcal{D}(A^n) \cap \mathcal{D}(A^{n+1})$ for any $n \in \mathbb{N}_0$, then $v \in \mathcal{D}\left(A^{\theta}\right)$ where $n \leq \theta \leq n+1$.

Proof. Considering part (i), let $y \in \ker A$ so that $Ay = 0$, and take $\theta > 0$. In the following, it will be shown that $y \in \ker A^{\theta}$. Indeed, consider the following spectral integrals, and let $\varepsilon > 0$

$$\begin{aligned}\left\|A^{\theta} y\right\|_{\mathcal{H}}^2 &= \int_{[0,\infty)} t^{2\theta}\, \mathrm{d}\, \langle y,\, \mathbf{E}(t)\, y\rangle \\ &= \int_{[0,\varepsilon)} t^{2\theta}\, \mathrm{d}\, \langle y,\, \mathbf{E}(t)\, y\rangle + \lim_{n\to\infty} \int_{[\varepsilon,n)} t^{2\theta}\, \mathrm{d}\, \langle y,\, \mathbf{E}(t)\, y\rangle \\ &\leq \varepsilon^{2\theta}\, \|y\|_{\mathcal{H}}^2 + \lim_{n\to\infty} n^{2\theta}\, \langle y,\, \mathbf{E}([\varepsilon, n))\, y\rangle\ ,\end{aligned}$$

where the Lebesgue monotone convergence theorem is used in passing to the second equality. Then, owing to the fact that $\|\mathbf{E}((0,\infty))\, y\|_{\mathcal{H}} = 0$ from

Lemma 6.2.4 (i), one has that due to the monotonicity of the scalar measure $\langle y,\, \mathbf{E}\,(t)\, y\rangle$, it follows $\langle y,\, \mathbf{E}\,([\varepsilon, n))\, y\rangle \leq \langle y,\, \mathbf{E}\,((0,\infty))\, y\rangle = 0$ for all $n \in \mathbb{N}$ and $\varepsilon > 0$. Putting this together,

$$\left\|A^{\theta} y\right\|_{\mathcal{H}}^{2} \leq \varepsilon^{2\theta} \langle y,\, \mathbf{E}\,([0,\varepsilon))\, y\rangle \leq \varepsilon^{2\theta} \left\|y\right\|_{\mathcal{H}}^{2} \xrightarrow{\varepsilon\to 0} 0\,.$$

Therefore $y \in \ker A^{\theta}$, and so it has been shown that $\ker A \subset \ker A^{\theta}$. For proving the opposite inclusion, let $w \in \ker A^{\theta}$. Now, by arguments similar to the proof of Lemma 6.2.4, it follows that as $t^{2\theta} > 0$ and is smooth on the set $(0,\infty)$, then if there exists *any* Borel set $\Omega \subset (0,\infty)$ such that $\langle w,\, \mathbf{E}\,(\Omega)\, w\rangle > 0$, then it *must* be that $\left\|A^{\theta} w\right\|_{\mathcal{H}}^{2} > 0$. This would contradict the fact that $w \in \ker A^{\theta}$. Therefore $\langle w,\, \mathbf{E}\,((0,\infty))\, w\rangle = 0$. By Lemma 6.2.4, it follows that $w \in \ker A$, so $\ker A^{\theta} \subset \ker A$ and part (i) is proven.

To prove part (ii), begin with $y \in \ker A$. Then, as $\ker A = \ker A^{\theta}$ by part (i), it follows that $A^{\theta} y = 0 \in \ker A$. Therefore $\ker A$ remains invariant under A^{θ}. To show the invariance of $(\ker A)^{\perp}$ under A^{θ}, let $y \in (\ker A)^{\perp} \cap \mathcal{D}\left(A^{\theta}\right)$. Then by Lemma 6.2.4 (ii) it follows that $y = \mathbf{E}\,((0,\infty))\, y$. So let $x \in \ker A = \ker A^{\theta}$ by part (i), from which

$$\left\langle x,\, A^{\theta} y\right\rangle = \left\langle A^{\theta} x,\, y\right\rangle = 0\,,$$

as A^{θ} is self-adjoint. Therefore, $A^{\theta} y \in (\ker A)^{\perp}$, and part (ii) is proven.

Working on part (iii), the cases $\theta = n$ or $\theta = n+1$ are trivial. First, however, consider the case for which $0 < \theta < 1$. Then $\left\|Av\right\|_{\mathcal{H}} < \infty$, so now

$$\begin{aligned}
\left\|A^{\theta} v\right\|_{\mathcal{H}}^{2} &= \int_{[0,\infty)} t^{2\theta}\, \mathrm{d}\,\langle v,\, \mathbf{E}\,(t)\, v\rangle \\
&\leq \int_{[0,1]} \mathrm{d}\,\langle v,\, \mathbf{E}\,(t)\, v\rangle + \int_{(1,\infty)} t^{2\theta}\, \mathrm{d}\,\langle v,\, \mathbf{E}\,(t)\, v\rangle \\
&\leq \left\|v\right\|_{\mathcal{H}}^{2} + \int_{(1,\infty)} t^{2\theta}\, \mathrm{d}\,\langle v,\, \mathbf{E}\,(t)\, v\rangle\ .
\end{aligned}$$

But on the interval $(1,\infty)$, one has that $t^{2} \geq t^{2\theta}$ as $\theta < 1$, so from the above it follows

$$\left\|A^{\theta}v\right\|_{\mathcal{H}}^{2} \leq \|v\|_{\mathcal{H}}^{2} + \int_{(1,\infty)} t^{2}\,\mathrm{d}\left\langle v,\,\mathbf{E}\left(t\right)v\right\rangle \leq \|v\|_{\mathcal{H}}^{2} + \|Av\|_{\mathcal{H}}^{2} < \infty\,.$$

Therefore $v \in \mathcal{D}\left(A^{\theta}\right)$ for the case of $0 \leq \theta \leq 1$. Now consider the situation $n < \theta < n+1$ for arbitrary $n \in \mathbb{N}$. An argument similar to the above may be repeated, but alternatively, the interpolation inequality [10, Chapter 4, Remark 2] may be used instead. Indeed, letting $\mu_v\left(t\right) = \left\langle v,\,\mathbf{E}\left(t\right)v\right\rangle$, one has that the condition $v \in \mathcal{D}\left(A^{n}\right) \cap \mathcal{D}\left(A^{n+1}\right)$ may be restated as $t \in L^{2n}([0,\infty),\mu_v\left(t\right)) \cap L^{2(n+1)}([0,\infty),\mu_v\left(t\right))$, and from the interpolation inequality, it follows that

$$\left\|A^{\theta}v\right\|_{\mathcal{H}}^{\frac{1}{\theta}} = \left(\int_{[0,\infty)} t^{2\theta}\,\mathrm{d}\mu_v\left(t\right)\right)^{\frac{1}{2\theta}} \leq \|t\|_{L^{2n}}^{\alpha}\,\|t\|_{L^{2(n+1)}}^{1-\alpha} = \|A^{n}v\|_{\mathcal{H}}^{\frac{\alpha}{n}}\left\|A^{n+1}v\right\|_{\mathcal{H}}^{\frac{1-\alpha}{n+1}}$$

for some $0 \leq \alpha \leq 1$. This completes the proof of part (iii). □

Lemma 6.2.6. *Let $z \in \mathcal{H}$. For a point $y \in \mathcal{S}$ as defined in Definition 6.2.2, the following conditions are equivalent.*

(i) $y = \mathcal{P}_{\mathcal{S}}z$,

(ii) $z - y \in (\ker A)^{\perp}$.

Proof. By the linearity of A, one has $\mathcal{S} = \{y\} + \ker A$. If $z - y \in (\ker A)^{\perp}$, then for any $x \in \ker A$, and hence a generic point $y + x \in \mathcal{S}$, it follows that

$$\|z - (y + x)\|_{\mathcal{H}}^{2} = \|z - y\|_{\mathcal{H}}^{2} + \|x\|_{\mathcal{H}}^{2} \geq \|z - y\|_{\mathcal{H}}^{2}\,,$$

so therefore y is necessarily the closest to z among all points in $\mathcal{S}$ (i.e., $y = \mathcal{P}_{\mathcal{S}}z$), showing (ii) ⇒ (i).

Conversely, if $y = \mathcal{P}_{\mathcal{S}}z$, and suppose by contradiction that $z - y$ does not belong to $(\ker A)^{\perp}$. Then $\mathfrak{Re}\left\langle x_0,\,z - y\right\rangle \neq 0$ for some $x_0 \in \ker A$, where $\mathfrak{Re}$ denotes the real part. In this case, consider the polynomial

$$p(t) := \|z - y - tx_0\|_{\mathcal{H}}^{2} = \|x_0\|_{\mathcal{H}}^{2}t^{2} - 2\mathfrak{Re}\left\langle x_0,\,z - y\right\rangle t + \|z - y\|_{\mathcal{H}}^{2}\,.$$

Clearly, depending on the sign of $\mathfrak{Re}\,\langle x_0,\, z-y\rangle$, one may choose $t\neq 0$ but small enough such that $p(t)\leq p(0)$. This shows that there exist points $y+tx_0\in\mathcal{S}$ for which $\|z-(y+tx_0)\|_{\mathcal{H}}\leq\|z-y\|_{\mathcal{H}}$, contradicting the assumption that y is the closest point to z among all points in $\mathcal{S}$. Then, necessarily, $z-y\in(\ker A)^{\perp}$, showing (i) ⇒ (ii). □

Now the definition of the iteration procedure that corresponds to a general conjugate gradient style method is presented.

Definition 6.2.7. Let $A:\mathcal{H}\to\mathcal{H}$ be a self-adjoint, positive operator in $\mathcal{H}$. Then given some $g\in\operatorname{ran}A$ and an initial guess $f^{[0]}$, for some $\theta\geq 0$ one defines the θ-iterates

$$f^{[N]}:=\underset{h\in\{f^{[0]}\}+\mathcal{K}_N(A,\,\mathfrak{R}_0)}{\operatorname{argmin}}\left\|A^{\theta/2}(h-\mathcal{P}_{\mathcal{S}}h)\right\|_{\mathcal{H}}\,, \tag{6.7}$$

provided that $f^{[0]}\in\mathcal{C}^{\infty}(A)$, $g\in\mathcal{C}^{\infty}(A)$. The residual $\mathfrak{R}_N$ at step N is defined as

$$\mathfrak{R}_N:=Af^{[N]}-g\,. \tag{6.8}$$

Remark 6.2.8. If $f^{[0]},g\in\mathcal{C}^{\infty}(A)$, then the method is well-defined for all $\theta\geq 0$, as $\mathcal{C}^{\infty}(A)\subset\mathcal{D}\left(A^{\theta/2}\right)$. This fact is due to Lemma 6.2.5 (iii). As a consequence of $f^{[0]},g\in\mathcal{C}^{\infty}(A)$ one has that $\mathfrak{R}_0\in\mathcal{C}^{\infty}(A)$ as well as $\mathcal{P}_{\mathcal{S}}h\in\mathcal{S}\subset\mathcal{C}^{\infty}(A)$, and so $\mathcal{K}(A,\,\mathfrak{R}_0)\subset\mathcal{C}^{\infty}(A)$.

Yet, given some $N\in\mathbb{N}$, the iterate $f^{[N]}$ is still well-defined under the weaker assumptions $f^{[0]}\in\mathcal{C}^{N}(A)$, $g\in\mathcal{C}^{N-1}(A)$ and $0\leq\theta\leq N-1$.

Only the setting $f^{[0]}\in\mathcal{C}^{\infty}(A)$ with $g\in\operatorname{ran}A\cap\mathcal{C}^{\infty}(A)$ shall be considered within this Chapter to ensure that the *infinite-dimensional* Krylov subspace $\mathcal{K}(A,\,\mathfrak{R}_0)$ is properly defined, and the θ-iterates are properly defined by (6.7).

From analogous considerations in Nemirovskiy and Polyak [64], the class of vectors $\mathfrak{C}_{A,g}(\theta)$ is introduced.

Definition 6.2.9. Let $\theta\in\mathbb{R}$, and $A:\mathcal{H}\to\mathcal{H}$ be a self-adjoint, positive operator in $\mathcal{H}$, with $g\in\operatorname{ran}A\cap\mathcal{C}^{\infty}(A)$. Then the class of vectors $\mathfrak{C}_{A,g}(\theta)$ is defined as

$$\mathfrak{C}_{A,g}\left(\theta\right) := \begin{cases} \{x \in \mathcal{H} \,|\, x - \mathcal{P}_{\mathcal{S}}x \in \mathcal{D}\left(A^{\theta/2}\right)\}\,, & \theta \geq 0 \\ \{x \in \mathcal{H} \,|\, x - \mathcal{P}_{\mathcal{S}}x \in \operatorname{ran}\left(A^{-\theta/2}\right)\}\,, & \theta < 0\,. \end{cases} \tag{6.9}$$

The dependence of $\mathfrak{C}_{A,g}\left(\theta\right)$ on the vector g is implicit through the solution manifold $\mathcal{S}$.

Remark 6.2.10. Distinguishing the two cases in (6.9) is necessary in the case that A has a non-trivial kernel, which makes the operator $A^{\theta/2}$ undefined when $\theta < 0$.

In the case where one has A as injective, it is permissible to define $A^{\theta/2}$ for strictly negative θ as A is invertible on its range ran$A$ which is dense in $\mathcal{H}$. As t is non-zero $\mathbf{E}$-a.e. in $\mathbb{R}$, Proposition B.2.18 applies, and additionally one has that $\mathcal{D}\left(A^{\theta/2}\right) = \operatorname{ran}(A^{-\theta/2})$.

Still following analogous discussions from [64], other useful notions are defined that relate to the convergence, and will become critical in the proofs that follow.

Definition 6.2.11. Let $\theta \in \mathbb{R}$ be fixed, and $x \in \mathfrak{C}_{A,g}\left(\theta\right)$ for A and g as given in Definition 6.2.9. Then the vector $u_\theta(x)$ is defined as

$$u_\theta(x) := \begin{cases} A^{\theta/2}(x - \mathcal{P}_{\mathcal{S}}x)\,, & \theta \geq 0 \\ \text{the minimal norm solution } u \text{ to } A^{-\theta/2}u = x - \mathcal{P}_{\mathcal{S}}x\,, & \theta < 0\,. \end{cases} \tag{6.10}$$

The functional $\rho_\theta(x)$ is defined on the vectors $x \in \mathfrak{C}_{A,g}\left(\theta\right)$ as

$$\rho_\theta(x) := \left\|u_\theta(x)\right\|^2_{\mathcal{H}}\,. \tag{6.11}$$

Remark 6.2.12. The definition of $\rho_\theta(x)$ facilitates a more concrete notion regarding its representation. As $\ker A$ and $(\ker A)^\perp$ remain invariant under the action of positive powers of A, one may write

$$\rho_\theta(x) = \begin{cases} \left\| A^{\theta/2}(x - \mathcal{P}_{\mathcal{S}}x) \right\|_{\mathcal{H}}^2 , & \theta \geq 0 \\ \left\| \left(A^{-\theta/2}\big|_{\mathrm{ran}(A^{-\theta/2})} \right)^{-1} (x - \mathcal{P}_{\mathcal{S}}x) \right\|_{\mathcal{H}}^2 , & \theta < 0 . \end{cases} \tag{6.12}$$

In this formulation for $\theta < 0$, the operator $A^{-\theta/2}\big|_{\mathrm{ran}(A^{-\theta/2})}$ is understood for a self-adjoint, positive-definite operator on the Hilbert subspace $\overline{\mathrm{ran}A} = \overline{\mathrm{ran}(A^{-\theta/2})}$. The fact that it remains self-adjoint is a consequence of the fact that $\overline{\mathrm{ran}A} = (\ker A)^\perp$ and both $\ker A$ and $(\ker A)^\perp$ remain *invariant* under A and its positive powers.

Proposition 6.2.13. *The θ-iterates $f^{[N]}$ defined for a given $\theta \geq 0$ by means of Definition 6.2.7 under the assumption $f^{[0]} \in \mathcal{C}^\infty(A)$, and $g \in \mathrm{ran}A \cap \mathcal{C}^\infty(A)$ satisfy the following.*

(i) $f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[N]} \in (\ker A)^\perp$ *for all* $N \in \mathbb{N}_0$,

(ii) $\mathcal{P}_{\mathcal{S}} f^{[N]} = \mathcal{P}_{\mathcal{S}} f^{[0]}$ *for all* $N \in \mathbb{N}$,

(iii) $f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[N]} = p_N(A)(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]})$ *for all* $N \in \mathbb{N}$,

where $p_N(t)$ is, for each N, a polynomial of degree up to N and such that $p_N(0) = 1$.

Proof. In the minimisation (6.7)

$$h - f^{[0]} = q_{N-1}(A)(Af^{[0]} - g) = q_{N-1}(A)A(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]})$$

for some polynomial $q_{N-1} \in \mathbb{P}_{N-1}$. From this, also

$$h - \mathcal{P}_{\mathcal{S}} f^{[0]} = q_{N-1}(A)A(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]}) + (f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]}) .$$

This implies, upon setting $p_N(t) := tq_{N-1}(t) + 1$, that

$$f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]} = p_N(A)(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]}) \quad \forall N \in \mathbb{N} , \tag{*}$$

where $p_N \in \mathbb{P}_N^{(1)}$.

Moreover, $f^{[N]} - \mathcal{P}_\mathcal{S} f^{[N]} \in (\ker A)^\perp$ as a consequence of Lemma 6.2.6 applied to $z = f^{[N]}$ and $y = \mathcal{P}_\mathcal{S} f^{[N]}$. With an analogous argument, $f^{[0]} - \mathcal{P}_\mathcal{S} f^{[0]} \in (\ker A)^\perp$, and so (i) is proved.

From part (i) and $f^{[0]} \in \mathcal{C}^\infty(A)$ with $g \in \mathrm{ran} A \cap \mathcal{C}^\infty(A)$, one has $f^{[0]} - \mathcal{P}_\mathcal{S} f^{[0]} \in (\ker A)^\perp \cap \mathcal{C}^\infty(A)$. Now, $(\ker A)^\perp \cap \mathcal{C}^\infty(A)$ is invariant under the action of polynomials of A (Lemma 6.2.5), and so from part (i) and (*), one has that $f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]} \in (\ker A)^\perp$.

Next, one may split as follows

$$\mathcal{P}_\mathcal{S} f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]} = (f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]}) - (f^{[N]} - \mathcal{P}_\mathcal{S} f^{[N]}) .$$

Obviously, $\mathcal{P}_\mathcal{S} f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]} \in \ker A$. But on the right hand side of the above, as just shown, one has both $f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]} \in (\ker A)^\perp$ and $f^{[N]} - \mathcal{P}_\mathcal{S} f^{[N]} \in (\ker A)^\perp$. As a consequence, $\mathcal{P}_\mathcal{S} f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]} \in (\ker A)^\perp \cap \ker A$ so that $\mathcal{P}_\mathcal{S} f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]} = 0$. This proves part (ii).

From (ii), (*) takes on the required form for part (iii). □

By (ii) in Proposition 6.2.13, all $f^{[N]}$'s have the same projection onto the solution manifold $\mathcal{S}$, so that the approach of $f^{[N]}$ to $\mathcal{S}$ is the same as the convergence $f^{[N]} \to \mathcal{P}_\mathcal{S} f^{[0]}$.

Lemma 6.2.14. *If $A \in \mathscr{B}(\mathcal{H})$ is self-adjoint and positive, and if $g \in \mathrm{ran} A$, then*

(i) $\mathfrak{C}_{A,g}(\theta) = \mathcal{H}$ *whenever* $\theta \geq 0$,

(ii) $\mathfrak{C}_{A,g}(\theta) \subset \mathfrak{C}_{A,g}(\theta')$ *for* $\theta \leq \theta'$,

(iii) for $\theta \leq \theta'$ *and* $x \in \mathfrak{C}_{A,g}(\theta)$, *one has* $u_{\theta'}(x) = A^{(\theta'-\theta)/2} u_\theta(x)$, *from which* $\rho_{\theta'}(x) \leq \|A\|_{\mathrm{op}}^{\theta'-\theta} \rho_\theta(x)$.

Proof. Part (i) is evident from the fact that $\mathcal{D}\left(A^{\theta/2}\right) = \mathcal{H}$ for any $\theta \geq 0$ as $A \in \mathscr{B}(\mathcal{H})$ and positive.

Part (ii) is obvious when $\theta' \geq 0$. If, instead, $\theta \leq \theta' < 0$, then $\mathrm{ran}(A^{-\theta/2}) \subset \mathrm{ran}(A^{-\theta'/2})$ owing to the boundedness and positivity of A. Indeed, let $v \in \mathrm{ran}(A^{-\theta/2})$, then there exists some $u \in \mathcal{H}$ such that $v = A^{-\theta/2} u$. So now $v =$

$A^{-\theta'/2}A^{-(\theta-\theta')/2}u$, so that in fact $v \in \operatorname{ran}(A^{-\theta'/2})$ and therefore $\operatorname{ran}(A^{-\theta/2}) \subset \operatorname{ran}(A^{-\theta'/2})$. Therefore part (ii) is valid for all $\theta \leq \theta'$.

If $0 \leq \theta \leq \theta'$, then

$$u_{\theta'}(x) = A^{\theta'/2}(x - \mathcal{P}_{\mathcal{S}}x) = A^{(\theta'-\theta)/2}A^{\theta/2}(x - \mathcal{P}_{\mathcal{S}}x) = A^{(\theta'-\theta)/2}u_{\theta}(x)\,.$$

If instead $\theta < 0 \leq \theta'$, then $u_{\theta'}(x) = A^{\theta'/2}(x - \mathcal{P}_{\mathcal{S}}x)$ and $A^{-\theta/2}u_{\theta}(x) = x - \mathcal{P}_{\mathcal{S}}x$, from which

$$A^{(\theta'-\theta)/2}u_{\theta}(x) = A^{\theta'/2}(x - \mathcal{P}_{\mathcal{S}}x) = u_{\theta'}(x)\,.$$

Lastly, if $\theta \leq \theta' < 0$, then $A^{-\xi/2}u_{\xi}(x) = x - \mathcal{P}_{\mathcal{S}}x$ for both $\xi = \theta$ and $\xi = \theta'$, and so from

$$x - \mathcal{P}_{\mathcal{S}}x = A^{-\theta/2}u_{\theta}(x) = A^{-\theta'/2}A^{(\theta'-\theta)/2}u_{\theta}(x)$$

and $A^{-\theta'/2}u_{\theta'}(x) = x - \mathcal{P}_{\mathcal{S}}x$ one deduces that $u_{\theta'}(x) = A^{(\theta'-\theta)/2}u_{\theta}(x)$. As such, the identity is proved for all the possible cases. The inequality $\rho_{\theta'}(x) \leq \|A\|_{\mathrm{op}}^{(\theta'-\theta)/2}\rho_{\theta}(x)$ immediately follows from the boundedness of A. This completes the proof of part (iii). □

In the unbounded setting, in order to evaluate certain ρ_θ-functionals along the sequence of the $f^{[N]}$'s, some extra assumptions on the initial guess $f^{[0]}$ are necessary.

Lemma 6.2.15. *Consider the θ-iterates $f^{[N]}$ defined for a given $\theta \geq 0$ by means of Definition 6.2.7 under the assumption $g \in \operatorname{ran}A \cap \mathcal{C}^\infty(A)$. Then*

(i) $f^{[N]} \in \mathfrak{C}_{A,g}(\sigma)$ for all $\sigma \geq 0$,

(ii) $f^{[N]} \in \mathfrak{C}_{A,g}(\sigma)$ for any $\sigma < 0$ such that, additionally $f^{[0]} \in \mathfrak{C}_{A,g}(\sigma)$, in which case

$$u_\sigma(f^{[N]}) = p_N(A)u_\sigma(f^{[0]})\,, \tag{6.13}$$

where $p_N(t)$ is the polynomial described in Proposition 6.2.13.

Proof. Proposition 6.2.13 (iii) and the fact that $g, f^{[0]} \in \mathcal{C}^{\infty}(A)$ combine with the interpolation result from Lemma 6.2.5 (iii), to give one that $f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[N]} \in \mathcal{D}\left(A^{\sigma/2}\right)$ for all $\sigma \geq 0$. This proves part (i).

Now assume that $f^{[0]} \in \mathfrak{C}_{A,g}(\sigma)$ for some $\sigma < 0$. In this case, Proposition 6.2.13 (iii) reads

$$f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[N]} = p_N(A)(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]}) = p_N(A) A^{-\sigma/2} u_{\sigma}(f^{[0]}) \,,$$

due to the definition (6.10) of $u_{\sigma}(f^{[0]})$. And so, due to the commutativity of polynomials of A with $A^{-\sigma/2}$ for the vector $u_{\sigma}(f^{[0]})$ (an application of Theorem B.2.16), it is clear that $f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[N]} \in \mathrm{ran} A^{-\sigma/2}$ and, again by (6.10), it follows that $u_{\sigma}(f^{[N]}) = p_N(A) u_{\sigma}(f^{[0]})$. This completes the proof of part (ii). □

The indicator ρ_{σ} is going to be used as the suitable *indicator of the convergence*. For the practical purposes, the most typical and meaningful choices for $\rho_{\sigma}(f^{[N]})$ are the following

$$\begin{aligned} \rho_0(f^{[N]}) &= \left\| f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]} \right\|_{\mathcal{H}}^2 \\ \rho_1(f^{[N]}) &= \left\langle f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]},\, A(f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]}) \right\rangle \\ \rho_2(f^{[N]}) &= \left\| A(f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]}) \right\|_{\mathcal{H}}^2 \,, \end{aligned} \tag{6.14}$$

that are, respectively, the norm of the error, the so-called 'energy' (semi-)norm, and the norm of the residual.

6.3 Measure-theoretic results

In this Section, some further technical properties that will be used to prove the main result are shown.

The following technical results are measure-theoretic in nature, but necessary to establish the convergence in the main result of the next section. A special role is played by the following measure

$$\mathrm{d}\mu_\sigma\,(t) := \mathrm{d}\left\langle u_\sigma(f^{[0]}),\, \mathbf{E}\,(t)\, u_\sigma(f^{[0]})\right\rangle \tag{6.15}$$

defined under the assumption that $f^{[0]} \in \mathfrak{C}_{A,g}\,(\sigma)$ for a given $\sigma \in \mathbb{R}$, and $u_\sigma(f^{[0]})$ as defined in (6.10). Clearly, by definition, μ_σ is a finite measure with

$$\mu_\sigma\,([0,\infty)) = \int_{[0,\infty)} \mathrm{d}\mu_\sigma\,(t) = \left\|u_\sigma(f^{[0]})\right\|^2_{\mathcal{H}}\,. \tag{6.16}$$

Two relevant properties of μ_σ follow.

Proposition 6.3.1. *For the given self-adjoint, positive operator A on $\mathcal{H}$, and for given $g \in \mathcal{C}^\infty\,(A)$, $\sigma \in \mathbb{R}$, $f^{[0]} \in \mathcal{C}^\infty\,(A) \cap \mathfrak{C}_{A,g}\,(\sigma)$, consider the measure μ_σ defined by* (6.15)*. Then one has*

(i)

$$\mathrm{d}\mu_\sigma\,(t) = t^\sigma \mathrm{d}\left\langle f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]},\, \mathbf{E}\,(t)\,(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]})\right\rangle\,, \tag{6.17}$$

(ii) the spectral value $t = 0$ is not an atom for μ_σ, i.e.,

$$\mu_\sigma\,(\{0\}) = 0\,. \tag{6.18}$$

Proof. The identity (6.17) for $\sigma \geq 0$ follows immediately from the definition (6.15) and from the definition (6.10) of $u_\sigma(f^{[0]}) = A^{\sigma/2}(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]})$ owing to the fact that

$$\mathrm{d}\left\langle A^\alpha \psi,\, \mathbf{E}\,(t)\, A^\alpha \psi\right\rangle = \lambda^{2\alpha} \mathrm{d}\left\langle \psi,\, \mathbf{E}\,(t)\,\psi\right\rangle\,,$$

for all $\alpha \geq 0$ and $\psi \in \mathcal{D}\,(A^\alpha)$ (a result from Proposition B.2.15).

If instead, $\sigma < 0$, then consider the auxiliary measures

$$\mathrm{d}\widetilde{\mu}_\sigma(t) := t^{-\sigma}\mathrm{d}\mu_\sigma\,(t)\,, \quad \mathrm{d}\hat{\mu}_\sigma(t) := \mathrm{d}\left\langle f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]},\, \mathbf{E}\,(t)\,(f^{[0]} - \mathcal{P}_{\mathcal{S}} f^{[0]})\right\rangle\,.$$

On an arbitrary Borel subset $\Omega \subset [0,\infty)$ one then has

$$\begin{aligned}
\widetilde{\mu}_\sigma(\Omega) &= \int_\Omega t^{-\sigma}\,\mathrm{d}\mu_\sigma\,(t) \\
&= \int_{[0,\infty)} t^{-\sigma}\chi_\Omega\,\mathrm{d}\mu_\sigma\,(t) \\
&= \left\|A^{-\sigma/2}\mathbf{E}\,(\Omega)\,u_\sigma(f^{[0]})\right\|^2_{\mathcal{H}} \\
&= \left\|\mathbf{E}\,(\Omega)\,A^{-\sigma/2}u_\sigma(f^{[0]})\right\|^2_{\mathcal{H}} \\
&= \left\|\mathbf{E}\,(\Omega)\,(f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]})\right\|_{\mathcal{H}} = \int_\Omega \mathrm{d}\hat{\mu}_\sigma(\Omega)\,,
\end{aligned}$$

from using the definition (6.10) of $f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]} = A^{-\sigma/2}u_\sigma(f^{[0]})$, along with (B.9). On a more technical note, the proof of $\left\|A^{-\sigma/2}\mathbf{E}\,(\Omega)\,u_\sigma(f^{[0]})\right\|^2_{\mathcal{H}} = \left\|\mathbf{E}\,(\Omega)\,A^{-\sigma/2}u_\sigma(f^{[0]})\right\|^2_{\mathcal{H}}$ requires a trivial modification of the proof of (B.9), namely that of using a bounding sequence $(M_n \cap \Omega)_{n\in\mathbb{N}}$ for the set Ω. This shows that $\mathrm{d}\widetilde{\mu}_\sigma(t) = \mathrm{d}\hat{\mu}_\sigma(t)$, from which one has (6.17). Part (i) is proved.

Moving on to part (ii), recall from Lemma 6.2.5 that $f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]} \in (\ker A)^\perp$. Therefore, $\hat{\mu}_\sigma(\{0\}) = 0$ (Lemma 6.2.4 (i)). So, (6.17) implies also that $\mu_\sigma\,(\{0\}) = 0$. □

Now a further set of technical results may be presented, specifically concerning the polynomial p_N in Proposition 6.2.13 (iii) of the ξ-iterates $f^{[N]}$, that correspond to the actual minimisation (6.7).

Proposition 6.3.2. *For the given self-adjoint, positive operator A on $\mathcal{H}$, and for given $g \in \mathcal{C}^\infty(A)$, $\sigma \in \mathbb{R}$, $f^{[0]} \in \mathcal{C}^\infty(A) \cap \mathfrak{C}_{A,g}(\sigma)$ and $\xi \geq 0$, let $f^{[N]}$ be the N-th ξ-iterate defined by* (6.7) *with initial guess $f^{[0]}$ and parameter $\theta = \xi$, and let*

$$s_N := \operatorname*{argmin}_{p_N\in\mathbb{P}^{(1)}_N} \int_{[0,\infty)} t^\xi p_N^2(t)\,\mathrm{d}\left\langle f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]},\,\mathbf{E}\,(t)\,(f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]})\right\rangle \tag{6.19}$$

for each $N \in \mathbb{N}$. Then the following properties hold.

(i) One has

$$f^{[N]}-\mathcal{P}_{\mathcal{S}}f^{[N]} = s_N(A)(f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]}) \quad \forall N \in \mathbb{N}\,. \tag{6.20}$$

(ii) *The family* $(s_N)_{N\in\mathbb{N}}$ *is a set of orthogonal polynomials on* $[0,\infty)$ *with respect to the measure*

$$\begin{aligned}\mathrm{d}\nu_\xi(t) &:= t^{\xi-\sigma+1}\mathrm{d}\mu_\sigma(t)\\ &= t^{\xi+1}\mathrm{d}\left\langle f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]},\,\mathbf{E}(t)\,(f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]})\right\rangle\end{aligned}\tag{6.21}$$

and satisfying $\deg s_N = N$ *and* $s_N(0)=1$ *for all* $N\in\mathbb{N}$.

(iii) *One has*

$$\rho_\sigma(f^{[N]}) = \int_{[0,\infty)} s_N^2(t)\,\mathrm{d}\mu_\sigma(t)\quad \forall N\in\mathbb{N}\,.\tag{6.22}$$

Proof. Temporarily denote $\widetilde{s}_N\in\mathbb{P}_N^{(1)}$ the polynomial that qualifies the iterate $f^{[N]}$ in Proposition 6.2.13 (iii) by means of the minimisation (6.7) with $\theta=\xi$. Then

$$\begin{aligned}\operatorname*{argmin}_{h\in\{f^{[0]}\}+\mathcal{K}_N(A,\mathfrak{R}_0)}\left\|A^{\xi/2}(h-\mathcal{P}_{\mathcal{S}}h)\right\|_{\mathcal{H}}^2 &= \left\|A^{\xi/2}(f^{[N]}-\mathcal{P}_{\mathcal{S}}f^{[N]})\right\|_{\mathcal{H}}^2\\ &= \left\|A^{\xi/2}\widetilde{s}_N(A)(f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]})\right\|_{\mathcal{H}}^2\\ &= \int_{[0,\infty)} t^\xi\widetilde{s}_N^2(t)\,\mathrm{d}\left\langle f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]},\,\mathbf{E}(t)\,(f^{[0]}-\mathcal{P}_{\mathcal{S}}f^{[0]})\right\rangle\,.\end{aligned}$$

Comparing the above with (6.19) it is immediate that $\widetilde{s}_N$ must be the polynomial s_N. Therefore the equation in Proposition 6.2.13 (iii) takes the form (6.20). This proves (i).

By way of (6.17), equation (6.20) may be rewritten

$$s_N = \operatorname*{argmin}_{p_N\in\mathbb{P}_N^{(1)}}\int_{[0,\infty)} t^{\xi-\sigma}p_N^2(t)\,\mathrm{d}\mu_\sigma(t)\,.$$

The minimising property of s_N implies that

$$\begin{aligned}0 &= \frac{\mathrm{d}}{\mathrm{d}\varepsilon}\bigg|_{\varepsilon=0}\int_{[0,\infty)} t^{\xi-\sigma}(s_N(t)+\varepsilon t q_{N-1}(t))^2\,\mathrm{d}\mu_\sigma(t)\\ &= 2\int_{[0,\infty)} t^{\xi-\sigma+1}s_N(t)q_{N-1}(t)\,\mathrm{d}\mu_\sigma(t)\end{aligned}$$

for any $q_{N-1} \in \mathbb{P}_{N-1}$ (indeed $s_N + \varepsilon t q_{N-1} \in \mathbb{P}_N^{(1)}$). Therefore, owing to (6.21),

$$\int_{[0,\infty)} s_N(t) q_{N-1}(t) \,\mathrm{d}\nu_\xi(t) = 0 \quad \forall q_{N-1} \in \mathbb{P}_{N-1}\,.$$

As this condition is valid for all $N \in \mathbb{N}$, it is well known [19, 92, 52] that this amounts to saying that $(s_N)_{N\in\mathbb{N}}$ is a set of orthogonal polynomials on $[0,\infty)$ with respect to the measure $\mathrm{d}\nu_\xi$. The fact that $s_N(0) = 1$ has already been demonstrated in Proposition 6.2.13. Part (ii) is proved.

Now, if $\sigma \geq 0$, then Proposition 6.2.13 (ii), (6.12), (6.17), and (6.20) together yield

$$\begin{aligned}\rho_\sigma(f^{[N]}) &= \big\|A^{\sigma/2}(f^{[N]} - \mathcal{P}_\mathcal{S} f^{[N]})\big\|_\mathcal{H}^2 = \big\|A^{\sigma/2} s_N(A)(f^{[0]} - \mathcal{P}_\mathcal{S} f^{[0]})\big\|_\mathcal{H}^2 \\ &= \int_{[0,\infty)} s_N^2(t) \,\mathrm{d}\mu_\sigma\,(t)\ .\end{aligned}$$

If, instead, one has that $\sigma < 0$, then from (6.20), the identity (6.13) reads

$$u_\sigma(f^{[N]}) = s_N(A) u_\sigma(f^{[0]})\,.$$

The latter identity, together with (6.12) and (6.15), yeild

$$\rho_\sigma(f^{[N]}) = \big\|u_\sigma(f^{[N]})\big\|_\mathcal{H}^2 = \big\|s_N(A) u_\sigma(f^{[0]})\big\|_\mathcal{H}^2 = \int_{[0,\infty)} s_N^2(t) \,\mathrm{d}\mu_\sigma\,(t)\ .$$

In either case, (6.22) is established, and part (iii) is proven. □

Remark 6.3.3. The measure ν_ξ too is finite, with

$$\int_{[0,\infty)} \mathrm{d}\nu_\xi = \Big\|A^{\frac{\xi+1}{2}}(f^{[0]} - \mathcal{P}_\mathcal{S} f^{[0]})\Big\|_\mathcal{H}^2\,, \tag{6.23}$$

as is evident from (6.21). Actually, one could define ν_ξ for arbitrary $\xi \geq -1$, but the restriction $\xi \geq 0$ is kept because here ξ is the parameter required in the definition of the ξ-iterates, and so must not be negative.

Remark 6.3.4. There is an implicit dependence on ξ in each s_N, as is clear from (6.19), analogously to the fact that the $f^{[N]}$'s depend on the choice of the parameter ξ. For convenience, this dependence is omitted from the notation s_N.

The key fact to take away from Proposition 6.3.2 (iii) is that the control of the convergence of the $f^{[N]}$'s in the ρ_σ-sense is tantamount to monitoring a precise spectral integral. As such, one may make use of both the properties of the orthogonal polynomials s_N and of the measure ν_ξ together to ensure appropriate convergence.

For convenience, the notation $\hat{s}_N$ is used to denote the corresponding *monic* polynomial to s_N, i.e.,

$$\hat{s}_N(t) := \left(\frac{1}{N!} \frac{\mathrm{d}^N}{\mathrm{d}t^N}\bigg|_{t=0} \right) s_N(t) \,. \tag{6.24}$$

The following proposition uses some specialised, technical results from the theory of orthogonal polynomials to establish some properties on the measure ν_ξ.

Proposition 6.3.5. *Consider the set $(s_N)_{N\in\mathbb{N}}$ of orthogonal polynomials on $[0,\infty)$ with respect to the measure ν_ξ, as defined in* (6.19) *and* (6.21) *under the assumptions of Proposition 6.3.2.*

(i) For each $N \in \mathbb{N}$, either $s_N(t) = 0$ ν_ξ-almost everywhere, or s_N has exactly N simple zeros, all located in $(0,\infty)$.

Assume now that the s_N's are all non-vanishing with respect to the ν_ξ measure, and denote by $\lambda_k^{(N)}$ the k-th zero of s_N, ordering the zeros as

$$0 < \lambda_1^{(N)} < \lambda_2^{(N)} < \cdots < \lambda_N^{(N)} \,. \tag{6.25}$$

(ii) (Separation.) One has

$$\lambda_k^{(N+1)} < \lambda_k^{(N)} < \lambda_{k+1}^{(N+1)} \quad \forall k \in \{1,2,\dots,N-1\} \,, \tag{6.26}$$

i.e., the zeros of s_N and s_{N+1} mutually separate one another.

(iii) (Monotonicity.) For each $k \in \mathbb{N}$,

$$\begin{aligned} &(\lambda_k^{(N)})_{N=k}^{\infty} \text{ is a decreasing sequence,} \\ &(\lambda_{N-k+1}^{(N)})_{N=k}^{\infty} \text{ is an increasing sequence.} \end{aligned} \tag{6.27}$$

In particular, the limits

$$\lambda_1 := \lim_{N\to\infty} \lambda_1^{(N)}, \quad \lambda_\infty := \lim_{N\to\infty} \lambda_N^{(N)} \tag{6.28}$$

exist in $[0,\infty) \cup \{\infty\}$.

(iv) (Representation.) The measure ν_ξ is actually supported only in the so-called 'true interval of orthogonality' $[\lambda_1, \lambda_\infty]$, and λ_1 is not an atom for ν_ξ, namely

$$\nu_\xi(\{\lambda_1\}) = 0\,. \tag{6.29}$$

Here, and in the following, the symbol $[\lambda_1, \lambda_\infty]$ is understood as the closure of $(\lambda_1, \lambda_\infty)$

(v) (Orthogonality.) One has

$$\int_{[0,\lambda_1^{(N)})} s_N^2(t) \frac{\lambda_1^{(N)}}{\lambda_1^{(N)} - t}\, \mathrm{d}\nu_\xi(t) = \int_{[\lambda_1^{(N)},\infty)} s_N^2(t) \frac{\lambda_1^{(N)}}{t - \lambda_1^{(N)}}\, \mathrm{d}\nu_\xi(t) \tag{6.30}$$

for any $N \in \mathbb{N}$.

Proof. Part (i) is a standard fact from the theory of orthogonal polynomials (see, e.g., [92, Theorem 3.3.1] or [52, Theorem 5.2]), from the fact that the map

$$\mathbb{P}([0,\infty)) \ni p \mapsto \int_{[0,\infty)} p(t)\, \mathrm{d}\nu_\xi(t)$$

is a positive-definite functional on $\mathbb{P}([0,\infty))$.

Part (ii) is another standard fact from the theory of orthogonal polynomials (see, e.g., [92, Theorem 3.3.2] or [19, Theorem I.5.3]).

Part (iii), is an immediate corollary of part (ii).

For part (iv), recall [19, Definition I.5.2] that the true interval of orthogonality $[\lambda_1, \lambda_\infty]$ is the smallest closed interval containing all the zeros $\lambda_k^{(N)}$. Moreover from [19, Theorem II.3.1], there exists a measure η in $[0, \infty)$ supported only on $[\lambda_1, \lambda_\infty]$ such that the s_N's remain orthogonal with respect to η too and

$$\tau_k := \int_{[0,\infty)} t^k \,\mathrm{d}\nu_\xi(t) = \int_{[\lambda_1,\lambda_\infty)} t^k \,\mathrm{d}\eta(t)\,, \quad \forall k \in \mathbb{N}_0\,.$$

The η-measure is actually a Stieltjes measure associated with a bounded, non-decreasing function ψ obtained as a point-wise limit of a subsequence of $(\psi_N)_{N\in\mathbb{N}}$, where

$$\psi_N(t) := \begin{cases} 0\,, & t < \lambda_1^{(N)}\,, \\ A_1^{(N)} + \cdots + A_p^{(N)}\,, & t \in [\lambda_p^{(N)}, \lambda_{p+1}^{(N)}) \text{ for } p \in \{1, \dots, N-1\}\,, \\ \mu_0\,, & t \geq \lambda_N^{(N)} \end{cases}$$

and $A_1^{(N)}, \dots, A_N^{(N)}$ are positive numbers determined by the Gauss quadrature formula (see [19] for details)

$$\tau_k = \sum_{p=1}^{N} A_p^{(N)} (\lambda_p^{(N)})^k\,, \quad \forall k \in \{0, 1, \dots, 2N-1\}\,.$$

Therefore,

$$\eta(\{\lambda_1\}) = \psi(\lambda_1) - \lim_{t\to\lambda_1^-} \psi(t) = 0\,,$$

because by part (ii) and (iii) $\lambda_1 < \lambda_1^{(N)}$ for all $N \in \mathbb{N}$, from which

$$\psi(\lambda_1) = \lim_{N\to\infty} \psi_N(\lambda_1) = 0$$

and

$$\psi(t) = \lim_{N\to\infty} \psi_N(t) = 0$$

for $t < \lambda_1$.

Next, one sees that $\nu_\xi = \eta$, i.e., the Hamburger moment problem that guarantees that $(s_N)_{N\in\mathbb{N}}$ is an orthogonal system on $[0,\infty)$ is *uniquely* solved with the measure ν_ξ. This follows from the classical criterion [89, Theorem 2.9] for the uniqueness of the orthogonality measure (see [52, Theorem 8.3] for a more modern discussion). Such a measure is unique *if and only if* $w(z)=0$ for some $z\in\mathbb{C}$, where

$$w(z) := \left(\sum_{N\in\mathbb{N}} |\hat{s}_N(z)|^2\right)^{-1}$$

and $(\hat{s}_N(z))_{N\in\mathbb{N}}$ is the monic system obtained from $(s_N)_{N\in\mathbb{N}}$ (see (6.24)). This is precisely the case on choosing $z=-1$ for $w(-1)=0$, as owing to (6.24) and (6.25), $\hat{s}_n(t)=\prod_{k=1}^{N}(t-\lambda_k^{(N)})$, from which

$$\hat{s}_N^2(-1) = \prod_{k=1}^{N}(-1-\lambda_k^{(N)})^2 > 1\,.$$

This shows that $\nu_\xi=\eta$, thus proving that ν_ξ is supported only on $[\lambda_1,\lambda_\infty]$ with $\nu_\xi(\{\lambda_1\})=0$.

Part (v) follows from the identity

$$\int_{[0,\infty)} s_N(t)q_{N-1}(t)\,\mathrm{d}\nu_\xi(t) = 0 \quad \forall q_{N-1}\in\mathbb{P}_{N-1}\,,$$

which has already been considered in the proof of Proposition 6.3.2 as a consequence of the orthogonality of the s_N's, when the explicit choice

$$q_{N-1}(t) := \frac{\lambda_1^{(N)}s_N(t)}{\lambda_1^{(N)}-t}$$

is made. □

Remark 6.3.6. Analogously to Remark 6.3.4, there is an implicit dependence on ξ of all the zeros $\lambda_k^{(N)}$. For a more convenient notation, this dependence is omitted.

In view of Proposition 6.3.5 (i), when the s_N's are not identically zero, they may be explicitly represented as

$$s_N(t) = \prod_{k=1}^{N} \left(1 - \frac{t}{\lambda_k^{(N)}}\right), \quad \hat{s}_N(t) = \prod_{k=1}^{N} (t - \lambda_k^{(N)}) . \tag{6.31}$$

The integral (6.30) plays a major role in the proof of the main result, and therefore the next technical lemma is needed to construct appropriate bounds.

Lemma 6.3.7. *Consider the set $(s_N)_{N\in\mathbb{N}}$ of orthogonal polynomials on $[0,\infty)$ with respect to the measure ν_ξ, as defined in (6.19) and (6.21) under the assumptions of Proposition 6.3.2 and with the further restriction $\xi - \sigma + 1 \geq 0$. Assume that the s_N's are non-zero polynomials with respect to the measure ν_ξ. Then, for any $N \in \mathbb{N}$,*

$$\int_{[0,\lambda_1^{(N)})} s_N^2(t) \frac{\lambda_1^{(N)}}{\lambda_1^{(N)} - t} \, \mathrm{d}\nu_\xi(t) \leq \mu_\sigma\left([0,\lambda_1^{(N)})\right) \left(\frac{\xi - \sigma + 1}{\delta_N}\right)^{\xi-\sigma+1} , \tag{6.32}$$

where

$$\delta_N := \frac{1}{\lambda_1^{(N)}} + 2\sum_{k=2}^{N} \frac{1}{\lambda_k^{(N)}} . \tag{6.33}$$

Remark 6.3.8. The estimate (6.32) provides a (ξ,σ)-dependent bound on a quantity that is ξ-dependent only. This is only possible for a constrained range of σ, namely $\sigma \leq \xi + 1$.

Proof of Lemma 6.3.7. For each $N \in \mathbb{N}$ consider the function

$$\begin{aligned}
[0,\lambda_1^{(N)}] \ni t \mapsto a_N(t) &:= \frac{\lambda_1^{(N)} t^{\xi-\sigma+1} s_N^2(t)}{\lambda_1^{(N)} - t} \\
&= t^{\xi-\sigma+1}\left(1 - \frac{t}{\lambda_1^{(N)}}\right) \prod_{k=2}^{N} \left(1 - \frac{t}{\lambda_k^{(N)}}\right)^2
\end{aligned}$$

(where the representation (6.31) is used for s_N), which is non-negative, smooth, and such that $a_N(0) = a_N(\lambda_1^{(N)}) = 0$. Let $t_N^* \in (0,\lambda_1^{(N)})$ be the point of maximum for a_N. Then $a'_N(t_N^*) = 0$, which from a standard computation yields

$$\xi - \sigma + 1 \geq t_N^* \left(\frac{1}{\lambda_1^{(N)}} + 2 \sum_{k=2}^{N} \frac{1}{\lambda_k^{(N)}} \right) = t_N^* \delta_N \,,$$

from which also

$$t_N^* \leq \frac{\xi - \sigma + 1}{\delta_N} \,.$$

Moreover, $0 \leq 1 - t/\lambda_1^{(N)}$ for $t \in [0, \lambda_1^{(N)}]$ and for all $k \in \{1, \dots, N\}$, as $\lambda_1^{(N)}$ is the smallest zero of s_N. Also, from examination of (6.31), it is immediate that $s_N^2(t) \leq 1$ for $t \in [0, \lambda_1^{(N)}]$. Therefore,

$$a_N(t) \leq a_N(t_N^*) \leq (t_N^*)^{\xi-\sigma+1} \leq \left(\frac{\xi - \sigma + 1}{\delta_N} \right)^{\xi-\sigma+1} \,, \quad t \in [0, \lambda_1^{(N)}] \,.$$

One may then conclude that

$$\begin{aligned} \int_{[0,\lambda_1^{(N)})} s_N^2(t) \frac{\lambda_1^{(N)}}{\lambda_1^{(N)} - t} \, \mathrm{d}\nu_\xi(t) &= \int_{[0,\lambda_1^{(N)})} a_N(t) \, \mathrm{d}\mu_\sigma \, (t) \\ &\leq \mu_\sigma \left([0, \lambda_1^{(N)}) \right) \left(\frac{\xi - \sigma + 1}{\delta_N} \right)^{\xi-\sigma+1} \,, \end{aligned}$$

which completes the proof. □

6.4 Main convergence result

Having the all the previous technical ingredients, the main result of this Chapter may now be stated and proved.

Theorem 6.4.1. *Let A be a self-adjoint, positive operator on the Hilbert space $\mathcal{H}$ and let $g \in \mathrm{ran} \cap \mathcal{C}^\infty(A)$. Consider the conjugate-gradient algorithm associated with A and g where the initial guess vector $f^{[0]}$ satisfies*

$$f^{[0]} \in \mathcal{C}^\infty(A) \cap \mathfrak{C}_{A,g}(\sigma^*) \,, \quad \sigma^* = \min\{\sigma, 0\}$$

for a given $\sigma \in \mathbb{R}$, and where the iterates $f^{[N]}$, $N \in \mathbb{N}$, are constructed via

(6.7) *with parameter* $\theta = \xi \geq 0$ *under the condition* $\sigma \leq \xi$. *Then*

$$\lim_{N\to\infty} \rho_\sigma(f^{[N]}) = 0\,.$$

In other words, the convergence holds at a given 'A-regularity level' σ for ξ-iterates built with *equal or higher* 'A-regularity level' $\xi \geq \sigma$, and with an initial guess $f^{[0]}$ that is A-smooth if $\sigma \geq 0$, and additionally belongs to the class $\mathfrak{C}_{A,g}\,(\sigma)$ if $\sigma < 0$.

Remark 6.4.2. If, for a finite N, $\rho_\sigma(f^{(N)}) = 0$, then the very iterate $f^{(N)}$ *is* a solution to the linear problem $Af = g$, and one says that the algorithm 'has come to convergence' in a finite number (N) of steps. Indeed, $\rho_\sigma(f^{(N)}) = 0$ is the same as $A^{\frac{\sigma}{2}}(f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]}) = 0$ if $\sigma \geq 0$, i.e., $f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]} \in \ker A^{\frac{\sigma}{2}} = \ker A$; this, combined with $f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]} \in (\ker A)^\perp$ (see Proposition 6.2.13 above), implies that $f^{[N]} = \mathcal{P}_{\mathcal{S}} f^{[0]} \in \mathcal{S}$. On the other hand, $\rho_\sigma(f^{(N)}) = 0$ is the same as $u_\sigma(f^{[N]}) = 0$ with $A^{-\frac{\sigma}{2}} u_\sigma(f^{[N]}) = f^{[N]} - \mathcal{P}_{\mathcal{S}} f^{[0]}$ if $\sigma < 0$, from which again $f^{[N]} = \mathcal{P}_{\mathcal{S}} f^{[0]} \in \mathcal{S}$.

Proof of Theorem 6.4.1. Obviously in the following, the assumption is that none of the orthogonal polynomials s_N, as defined by (6.19), vanish with respect to the measure ν_ξ introduced in (6.21). Otherwise, by Remark 6.4.2, the conjugate-gradient algorithm has come to convergence in a finite number of steps. The conclusion of Theorem 6.4.1 is then trivially true.

From the relation (6.21) between the measures μ_σ and ν_ξ and the fact that the latter is supported only on the true interval of orthogonality $[\lambda_1, \lambda_\infty]$ with no atom at λ_1 (Proposition 6.3.5 (iv)), the measure μ_σ too is only supported on such an interval with $\mu_\sigma(\{\lambda_1\}) = 0$. Thus, in practise,

$$\rho_\sigma(f^{[N]}) = \int_{[\lambda_1,\lambda_\infty]} s_N^2(t)\,\mathrm{d}\mu_\sigma(t)\ .$$

It is convenient to split integrals as follows

$$\begin{aligned}\int_{[0,\infty)} s_N^2(t)\,\mathrm{d}\mu_\sigma\,(t) &= \int_{[0,\lambda_1^{(N)})} s_N^2(t)\,\mathrm{d}\mu_\sigma\,(t) + \int_{[\lambda_1^{(N)},\infty)} s_N^2(t)\,\mathrm{d}\mu_\sigma\,(t) \\ &\leq \mu_\sigma\left([0,\lambda_1^{(N)})\right) + \int_{[\lambda_1^{(N)},\infty)} s_N^2(t)\,\mathrm{d}\mu_\sigma\,(t)\ ,\end{aligned} \tag{6.34}$$

from the fact that $s_N^2(t) \leq 1$ for $t \in [0,\lambda_1^{(N)})$.

In what follows, it will be shown that

$$\int_{[\lambda_1^{(N)},\infty)} s_N^2(t)\,\mathrm{d}\mu_\sigma\,(t) \leq \frac{1}{(\lambda_1^{(N)})^{\xi-\sigma+1}} \int_{[0,\lambda_1^{(N)})} s_N^2(t)\frac{\lambda_1^{(N)}}{\lambda_1^{(N)}-t}\,\mathrm{d}\nu_\xi(t)\,. \tag{6.35}$$

Actually, (6.35) is a consequence of the properties of s_N already discussed. Indeed, consider the inequality

$$\begin{aligned}1 \leq \left(\frac{t}{\lambda_1^{(N)}}\right)^{\xi-\sigma} &= \frac{1}{(\lambda_1^{(N)})^{\xi-\sigma+1}}\cdot\frac{\lambda_1^{(N)}}{t}\cdot t^{\xi-\sigma+1} \\ &\leq \frac{1}{(\lambda_1^{(N)})^{\xi-\sigma+1}}\cdot\frac{\lambda_1^{(N)}}{t-\lambda_1^{(N)}}\cdot t^{\xi-\sigma+1} \quad (t \geq \lambda_1^{(N)})\,,\end{aligned} \tag{6.36}$$

which is valid owing to the constraint $\xi - \sigma \geq 0$. Then,

$$\begin{aligned}\int_{[\lambda_1^{(N)},\infty)} s_N^2(t)\,\mathrm{d}\mu_\sigma\,(t) &\leq \frac{1}{(\lambda_1^{(N)})^{\xi-\sigma+1}} \int_{[\lambda_1^{(N)},\infty)} s_N^2(t)\frac{\lambda_1^{(N)}}{t-\lambda_1^{(N)}}\,\mathrm{d}\nu_\xi(t) \\ &= \frac{1}{(\lambda_1^{(N)})^{\xi-\sigma+1}} \int_{[0,\lambda_1^{(N)})} s_N^2(t)\frac{\lambda_1^{(N)}}{\lambda_1^{(N)}-t}\,\mathrm{d}\nu_\xi(t)\,,\end{aligned}$$

having used (6.21) and (6.36) in the first step, and (6.30) in the second. The estimate (6.35) is now proved.

Now, plugging (6.35) into (6.34) and applying Lemma 6.3.7 yields

$$\begin{aligned}\rho_\sigma(f^{[N]}) &= \int_{[0,\infty)} s_N^2(t)\,\mathrm{d}\mu_\sigma\,(t) \\ &\leq \mu_\sigma\left([0,\lambda_1^{(N)})\right) + \frac{\mu_\sigma\left([0,\lambda_1^{(N)})\right)}{(\lambda_1^{(N)})^{\xi-\sigma+1}}\left(\frac{\xi-\sigma+1}{\delta_N}\right)^{\xi-\sigma+1}\,.\end{aligned} \tag{6.37}$$

The second summand on the right hand side of (6.37) is estimated as

$$\frac{\mu_\sigma\left([0,\lambda_1^{(N)})\right)}{(\lambda_1^{(N)})^{\xi-\sigma+1}}\left(\frac{\xi-\sigma+1}{\delta_N}\right)^{\xi-\sigma+1}\leq(\xi-\sigma+1)^{\xi-\sigma+1}\mu_\sigma\left([0,\lambda_1^{(N)})\right)\,, \tag{6.38}$$

due to the fact that $\lambda_1^{(N)}\delta_N\geq 1$ (as seen from (6.33)) and the exponent $\xi-\sigma+1$ is positive. Thus

$$\rho_\sigma(f^{[N]})\leq\left(1+(\xi-\sigma+1)^{\xi-\sigma+1}\right)\mu_\sigma\left([0,\lambda_1^{(N)})\right)\,. \tag{6.39}$$

Recalling that $\mu_\sigma\left([0,\lambda_1^{(N)})\right)=\mu_\sigma\left([\lambda_1,\lambda_1^{(N)})\right)$, clearly

$$\mu_\sigma\left([0,\lambda_1^{(N)})\right)\xrightarrow{N\to\infty}0$$

because $\lambda_1^{(N)}\to\lambda_1$ from above, and $\mu_\sigma\left(\{\lambda_1\}\right)=0$. This, incidentally, also covers the case when $\lambda_1=0$, as $\mu_\sigma\left(\{0\}\right)=0$ as already discussed.

As a consequence of the above, $\rho_\sigma(f^{(N)})\to 0$ as $N\to\infty$. □

Remark 6.4.3. In the special case that A is actually everywhere defined and bounded, then the A-smoothness assumption (i.e., $f^{[0]},g\in\mathcal{C}^\infty(A)$) is automatically satisfied. Therefore, one only need assume that $g\in\operatorname{ran}A$, i.e., that the problem $Af=g$ is actually *solvable*, along with $f^{[0]}\in\mathfrak{C}_{A,g}(\sigma^*)$ for some $\sigma\in\mathbb{R}$ (where $\sigma^*=\min\{\sigma,0\}$), for the convergence $\rho_\sigma(f^{[N]})\to 0$ of the ξ-iterates to hold ($\xi\geq\sigma$). Then, due to Lemma 6.2.14, one automatically has that $\rho_{\sigma'}(f^{[N]})\to 0$ for any $\sigma'\geq\sigma$. This is exactly what was originally stated by Nemirovskiy and Polyak [64].

Therefore, in the bounded case, if σ is the minimum level of convergence, then not only are the ξ-iterates with $\xi\geq\sigma$ proven to ρ_σ-converge, but in addition they also $\rho_{\sigma'}$-converge at any other level $\sigma'\geq\sigma$.

This is all in contrast to the unbounded case, where the above comments generally cannot be exported.

Remark 6.4.4. In retrospect, the assumption $\xi\geq\sigma$ was necessary to establish the bound (6.35), but more precisely, the inequality (6.36). In the other steps, namely in (6.37) (which is an application of Lemma 6.3.7) and (6.38), only the less restrictive assumption $\xi - \sigma + 1 \geq 0$ was needed.

Remark 6.4.5. Estimate (6.39) in the proof shows that the vanishing rate of $\rho_\sigma(f^{[N]})$ is actually controlled by the vanishing rate of $\mu_\sigma\left([\lambda_1, \lambda_1^{(N)})\right)$.

In the original work by Nemirovskiy and Polyak [64], the vanishing rate of $\rho_\theta(f^{[N]})$ was actually quantified for some $\theta > \sigma$ (c.f. Chapter 2, Theorem 2.3.6). Here, it is impossible for one to modify the polynomial min-max argument used in [64], as the original argument relies on the finiteness of the interval over which the orthogonal polynomials s_N have their zeros (i.e., it relies on the boundedness of $\sigma(A)$).

As such, for a general unbounded A, it is reasonable to expect that convergence rates for $\rho_\sigma(f^{[N]})$ may be arbitrarily slow. That is to say, that during the course of running the algorithm and monitoring convergence for a finite number of iterations, it may appear that the estimate $\rho_\sigma(f^{[N]})$ stagnates to some finite value above 0. In reality however, the vanishing of $\rho_\sigma(f^{[N]})$ is guaranteed in the *limit* $N \to \infty$.

Remark 6.4.6. In Nemirovskiy and Polyak [64], the authors provide, for the A-bounded case, an explicit convergence rate for any $\rho_\theta(f^{[N]})$ where $\theta \in (\sigma, \xi]$ based on the polynomial min-max argument as mentioned in Remark 6.4.5. The convergence rate presented in [64, Theorem 7] is

$$\rho_\theta(f^{[N]}) \lesssim (2N+1)^{-2(\theta-\sigma)}\rho_\sigma(f^{[0]}), \tag{6.40}$$

which, in addition, was proven to be the *optimal* rate for the class of positive, bounded operators on $\mathcal{H}$ [65].

As the arguments leading to this rate cannot be repeated for the unbounded case, it is reasonable to expect that this rate may be violated in the A-unbounded setting. Section 6.5 contains numerical evidence illustrating this phenomenon.

Remark 6.4.7. Where exactly the true interval of orthogonality lies within $[0, +\infty)$ depends on the behaviour of the zeroes of the s_N's. In particular, in terms of the quantity δ_N defined in (6.33), there are two alternative scenarios:

(i) Case I: $\delta_N \to \infty$ as $N \to \infty$;

(ii) Case II: δ_N remains uniformly bounded, strictly above 0, in N.

If the operator A is bounded, then Case I applies. Indeed, the orthogonal polynomials s_N are defined on $\sigma(A) \subset [0, \|A\|_{\mathrm{op}}]$, and their zeroes cannot exceed $\|A\|_{\mathrm{op}}$, forcing δ_N to blow up with N. Moreover, $\lambda_\infty = \lim_{N\to\infty} \lambda_N^{(N)} < \infty$.

If instead A is unbounded, the $\lambda_k^{(N)}$'s fall in $[0, \infty)$ and depending on their rate of possible accumulation at infinity δ_N may still diverge as $N \to \infty$ or stay bounded.

Clearly in Case II one has $\lambda_1 > 0$ and $\lambda_N = \infty$, for otherwise the condition $\lambda_1 = \lim_{N\to\infty} \lambda_1^{(N)} = 0$ or $\lambda_\infty = \lim_{N\to\infty} \lambda_N^{(N)} < \infty$ would necessarily imply $\delta_N \to \infty$. Thus, in Case II the true interval of orthogonality is $[\lambda_1, \infty)$ and it is separated from zero.

Remark 6.4.8. It is worth mentioning that the convergence phenomena explained in Theorem 6.4.1 for *the* conjugate gradient method, i.e., precisely the case that the $f^{[N]}$'s are the 1-iterates, guarantees convergence of the *error* term $\rho_0(f^{[N]}) = \left\|f^{[N]} - \mathcal{P}_\mathcal{S} f^{[0]}\right\|_\mathcal{H}^2$, however it may still happen that the *residual* term $\rho_2(f^{[N]}) = \left\|Af^{[N]} - g\right\|_\mathcal{H}^2 = \|\mathfrak{R}_N\|_\mathcal{H}^2$ diverges. Therefore in the more general setting that $\xi < 2$, a-priori one does not have convergence guaranteed in the graph norm of the operator A.

This lack of control on the graph norm convergence is from a combination of the subtle restriction that $\sigma \leq \xi$ in the proof of Theorem 6.4.1, along with the fact that in Lemma 6.2.14 point (iii) only holds when A is bounded. In other words, due to the possible unboundedness of A, one is unable to control the convergence in a higher regularity than $\sigma = \xi (< 2)$. Indeed, Section 6.5 contains numerical evidence of this phenomenon.

Of course then, to have a guaranteed control of the residual convergence one must pick $\xi \geq 2$. In the special case of $\xi = 2$, this corresponds to a *residual minimisation* scheme at each step N.

Remark 6.4.9. In the original proof from [64] the vanishing of the $\rho_\sigma(f^{[N]})$'s was guaranteed by the blowing up of the sum δ_N, along with the simultaneous vanishing of a cleverly chosen sequence $(\gamma_N)_{N\in\mathbb{N}}$, where $\gamma_N = \min\{\lambda_1^{(N)}, \delta_N^{1/2}\}$ for all $N \in \mathbb{N}$. In [64], instead of splitting the integral as in (6.34), the authors separate it into small and large spectral values at the threshold γ_N. After a somewhat lengthy analysis, Nemirovskiy and Polyak reduce both parts of the integral for $t < \lambda_1^{(N)}$ and $t \geq \lambda_1^{(N)}$ to one over $[0, \gamma_N)$. These facts, together with $\mu_\sigma(\{0\}) = 0$, were necessary to guarantee the vanishing of the $\rho_\sigma(f^{[N]})$'s.

On the other hand, the proof in this Chapter bypasses the use of γ_N by using the properties of the orthogonal polynomials generated by the general iteration method (6.7). More specifically,

(i) the uniqueness of the measure ν_ξ (and hence μ_σ),

(ii) the interval of support for μ_σ,

(iii) and the fact that μ_σ has no atom at λ_1,

all together control the convergence using continuity properties of measures. This allows one to *simultaneously* consider cases where δ_N is bounded or unbounded in the limit $N \to \infty$ and do away with the sequence $(\gamma_N)_{N\in\mathbb{N}}$, effectively making this proof independent of the behaviour of δ_N and γ_N. In the case that δ_N remains uniformly bounded above 0 (only possible when A is unbounded), the original argument in [64] cannot be suitably modified, and an argument as used in this work becomes necessary.

6.5 Numerical Tests and Examples

In this Section some basic numerical tests that illustrate the main features are discussed, particularly in contrast to the bounded case. Here, $\mathcal{H} = L^2(\mathbb{R})$ is the choice of Hilbert space. Throughout the numerical computations, symbolic packages are used to bypass the discretisation of the problem. There are four main tests covering two differential operators, and two solution functions $f(x)$, for the inverse linear problem $(Af)(x) = g(x)$, used to generate the datum $g(x)$.

1. $A = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \mathbb{1}$, $\mathcal{D}(A) = H^2(\mathbb{R})$

   (a) $f(x) = \exp(-x^2)$, $g(x) = (3 - 4x^2)\exp(-x^2)$

   (b) $f(x) = \frac{1}{1+x^2}$, $g(x) = \frac{1}{1+x^2} + \frac{2}{(1+x^2)^2} - \frac{8x^2}{(1+x^2)^3}$

2. $A = -\frac{\mathrm{d}^2}{\mathrm{d}x^2}$, $\mathcal{D}(A) = H^2(\mathbb{R})$

   (a) $f(x) = \exp(-x^2)$, $g(x) = (2 - 4x^2)\exp(-x^2)$

   (b) $f(x) = \frac{1}{1+x^2}$, $g(x) = \frac{2}{(1+x^2)^2} - \frac{8x^2}{(1+x^2)^3}$

where H^2 denotes the Sobolev space of the second order. The operator $A = -\frac{\mathrm{d}^2}{\mathrm{d}x^2} + \mathbb{1}$ has a bounded, everywhere defined inverse, while $A = -\frac{\mathrm{d}^2}{\mathrm{d}x^2}$ does not have a bounded inverse on its range. In both cases, A is positive definite (and thus injective), and $f \in H^2(\mathbb{R}) \cap \mathcal{C}^\infty(A)$ ensuring that $g \in \operatorname{ran} A \cap \mathcal{C}^\infty(A)$.

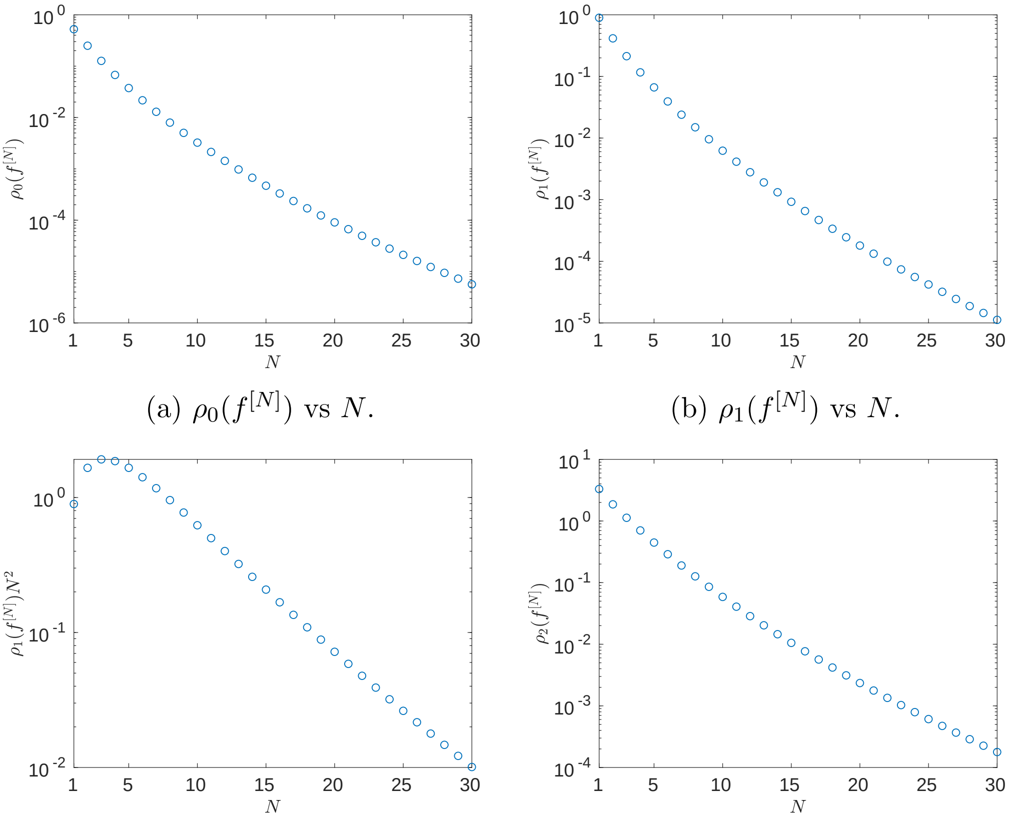
Numerical approximations to the linear inverse problem $Af = g$ are constructed using the initial guess $f^{[0]} = \mathbf{0}$, the zero function on $\mathbb{R}$, and the conjugate-gradient method, i.e., the ξ-iterates $f^{[N]}$ for the case $\xi = 1$. Each $f^{[N]}$ is in the Krylov subspace $\mathcal{K}_N(A, g) = \operatorname{span}\{g, Ag, \dots, A^{N-1}g\}$, and as g and $f^{[0]}$ are smooth, the 1-iterates $f^{[N]}$, and thus $\mathcal{K}(A, g)$, are indeed well-defined.

In practice, the minimisation (6.7) for $\theta = \xi = 1$ is implemented using the equivalent algebraic construction for the $f^{[N]}$'s [55, 87], along with symbolic computation packages.

The behaviour of the convergence is monitored using the three indicators (6.14) as N increases. Obviously $f^{[0]} \in \mathfrak{C}_{A,g}(\sigma)$ for all $\sigma \geq 0$, so that Theorem 6.4.1 guarantees that $\rho_\sigma(f^{[N]}) \to 0$ as $N \to \infty$ for any $\sigma \in [0, 1]$. In the bounded case, the residual $\rho_2(f^{[N]})$ would also be guaranteed to vanish (c.f., Lemma 6.2.14), yet this is not guaranteed a-priori in the unbounded setting.

Another meaningful quantity that is measured is $N^2\rho_1(f^{[N]})$. Indeed, if A was bounded the quantity $\rho_1(f^{[N]})$ is predicted to vanish *not slower than* N^{-2}, as may be seen in (6.40). Detecting the failure of $N^2\rho_1(f^{[N]})$ to remain uniformly bounded with N is a signature of the fact that one cannot apply the convergence rate analysis by Nemirovskiy and Polyak [64] for the A-bounded setting to the A-unbounded setting.

The results of the numerical tests are shown in Figures 6.1 to 6.4.



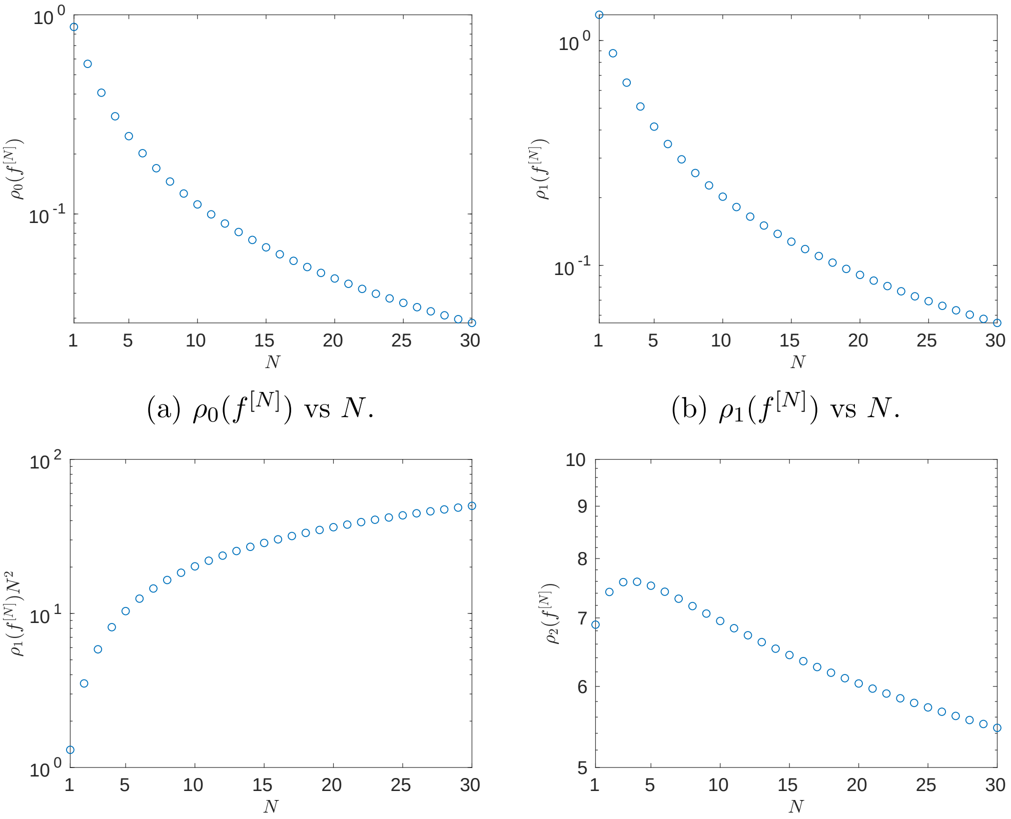
(a) $\rho_0(f^{[N]})$ vs N.

(b) $\rho_1(f^{[N]})$ vs N.

(c) $N^2\rho_1(f^{[N]})$ vs N.

(d) $\rho_2(f^{[N]})$ vs N.

Figure 6.1: Numerical tests for case 1 (a).



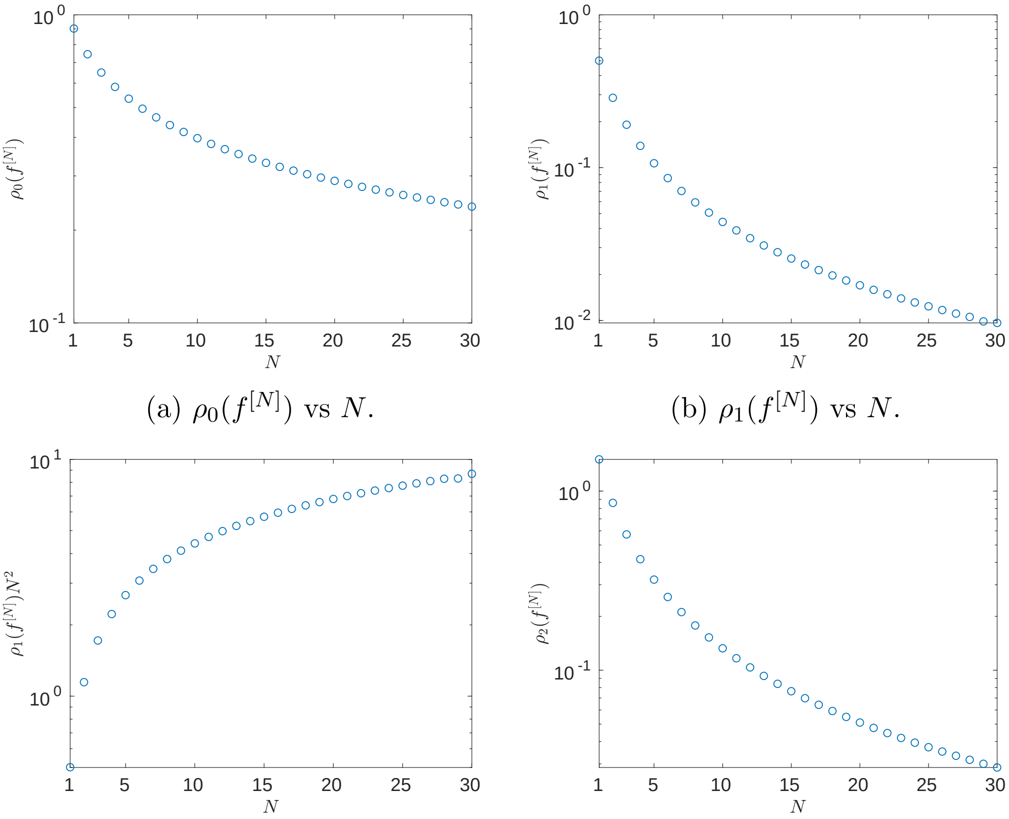
(a) $\rho_0(f^{[N]})$ vs N.

(b) $\rho_1(f^{[N]})$ vs N.

(c) $N^2\rho_1(f^{[N]})$ vs N.

(d) $\rho_2(f^{[N]})$ vs N.

Figure 6.2: Numerical tests for case 1 (b).



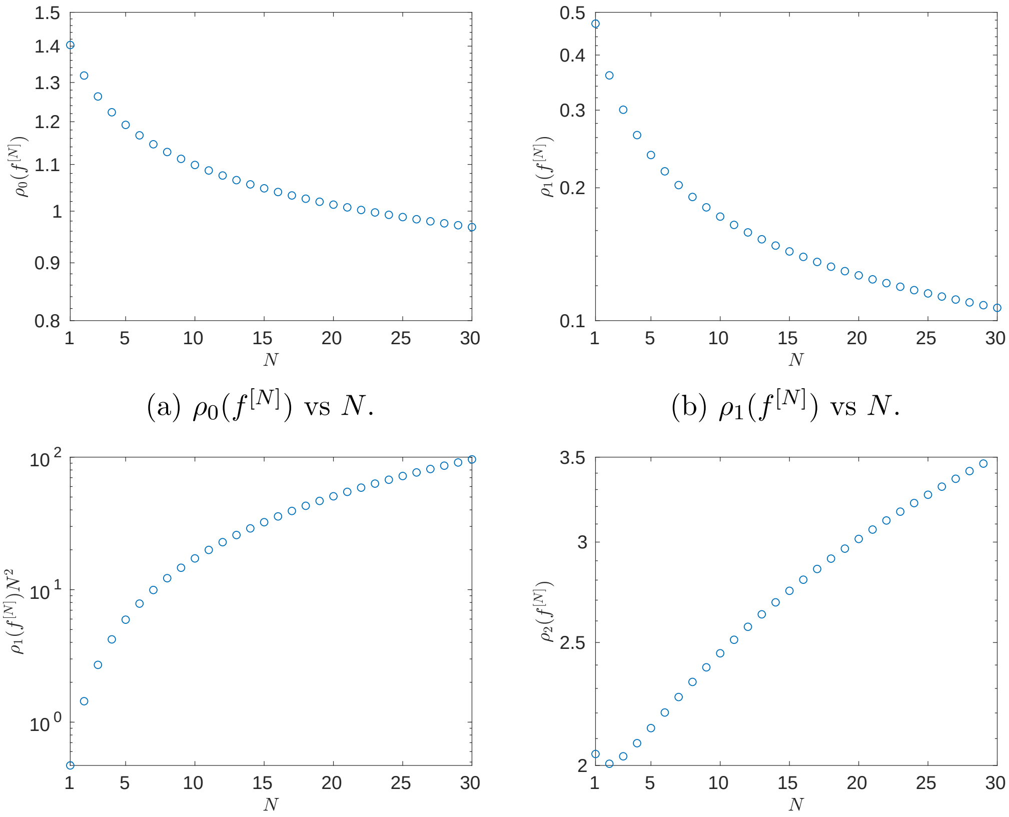
(a) $\rho_0(f^{[N]})$ vs N.

(b) $\rho_1(f^{[N]})$ vs N.

(c) $N^2\rho_1(f^{[N]})$ vs N.

(d) $\rho_2(f^{[N]})$ vs N.

Figure 6.3: Numerical tests for case 2 (a).



(a) $\rho_0(f^{[N]})$ vs N.

(b) $\rho_1(f^{[N]})$ vs N.

(c) $N^2\rho_1(f^{[N]})$ vs N.

(d) $\rho_2(f^{[N]})$ vs N.

Figure 6.4: Numerical tests for case 2 (b).

Case 1 (a) (Figure 6.1) reveals that the iterates converge in the sense of the error $\rho_0(f^{[N]})$ and the energy norm error $\rho_1(f^{[N]})$, as predicted by the theory, but also in the residual sense $\rho_2(f^{[N]})$. Moreover, the classical Nemirovskiy and Polyak convergence rate for $\rho_1(f^{[N]})$ is not violated, as $N^2\rho_1(f^{[N]})$ clearly decays.

In Case 2 (a) (Figure 6.3), the invertibility of A over all of $\mathcal{H}$ is lost, and here one sees that all the indicators of convergence, $\rho_0(f^{[N]})$, $\rho_1(f^{[N]})$, and $\rho_2(f^{[N]})$ all approach zero as N increases. Yet, the Nemirovskiy and Polyak convergence rate is violated, as $N^2\rho_1(f^{[N]})$ clearly increases with N.

In contrast to Cases 1 (a) and 2 (a), Cases 1 (b) and 2 (b) are performed using a function $f(x)$ that does not decay as a Gaussian, but instead has a long tail at large x. This feature now affects the convergence at higher regularity levels. More precisely, Case 1 (b) (Figure 6.2) shows the vanishing of $\rho_0(f^{[N]})$, $\rho_1(f^{[N]})$, and $\rho_2(f^{[N]})$, but unlike Case 1 (a), it now violates the Nemirovskiy and Polyak convergence rate for $\rho_1(f^{[N]})$ as one clearly sees $N^2\rho_1(f^{[N]})$ increasing with N.

Case 2 (b) (Figure 6.4) again confirms the theory predicted by Theorem 6.4.1 in that $\rho_0(f^{[N]})$ and $\rho_1(f^{[N]})$ decay with N to zero, but again, the latter does not follow the Nemirovskiy and Polyak convergence rate as $N^2\rho_1(f^{[N]})$ is clearly not uniformly bounded with N. More seriously in this case, the residual $\rho_2(f^{[N]})$ clearly *fails to converge.*

Chapter 7

Future Perspectives

7.1 Two remaining questions

From the previous chapters there still are some remaining questions, mostly regarding concrete examples of theoretical properties already discussed. More specifically, still, it remains

1. to produce an operator $A : \mathcal{H} \to \mathcal{H}$, injective, densely defined, closed and unbounded, with a $g \in \mathcal{H}$ such that given $f \in \mathcal{D}(A)$ with $Af = g$, one also has $P_{\mathcal{K}} f \in \mathcal{D}(A)$. Recall that $P_{\mathcal{K}}$ is the orthogonal projection onto the space $\overline{\mathcal{K}(A, g)}$. Having $P_{\mathcal{K}} f \in \mathcal{D}(A)$ is a requirement of Proposition 5.4.4.

   Clearly though, by the conjugate-gradient analysis of Chapter 6, some $A = A^* \geq 0$ with $g \in \operatorname{ran} A \cap \mathcal{C}^{\infty}(A)$ does the job,

2. to produce some operator $A = A^*$ such that $\mathcal{K}(A, g)$ is not a core for $A|_{\overline{\mathcal{K}(A, g)} \cap \mathcal{D}(A)}$. And, additionally, to have A such that it is *not* Krylov reducible.

7.2 Krylov methods and perturbed spectra

An intriguing area to investigate is the notion of Krylov solvability under perturbations of the operator A (which may also be cast as a perturbation in the datum g), and under what conditions Krylov-solvability 'survives' in the vanishing limit of the perturbation. To be a bit more clear, this is the investigation of the linear inverse problem on Hilbert space $\mathcal{H}$

$$A_\varepsilon f_\varepsilon = g\,, \tag{7.1}$$

where A_ε is a perturbation in $\|\cdot\|_{\mathrm{op}}$ of a known operator $A \in \mathscr{B}(\mathcal{H})$, g is the datum, and f_ε is the solution to the perturbed equation (7.1) (if one exists).

Generally speaking, perturbations of a linear operator can have very wild effects on the spectral properties. However, it is known that for the class of operators $\mathscr{B}(\mathcal{H})$, the spectrum is upper semicontinuous under a perturbation in the operator norm [51, Chapter IV, Theorem 3.1]. This leads to the following proposition.

Proposition 7.2.1. *Let $A \in \mathscr{B}(\mathcal{H})$ be a class-$\mathscr{K}$ operator, as described in Theorem 4.4.6, and consider the linear inverse problem $Af = g$, for $g \in \mathcal{H}$. Then there exists some $\varepsilon_0 > 0$ such that for all $A_\varepsilon \in \mathscr{B}(\mathcal{H})$ with $\|A - A_\varepsilon\|_{\mathrm{op}} < \varepsilon_0$, the linear inverse problem $A_\varepsilon f_\varepsilon = g$ is Krylov solvable.*

In particular, the Krylov solvability of the perturbed linear inverse problem $A_\varepsilon f_\varepsilon = g$ survives in the limit $\varepsilon \to 0$.

Proof. Consider the operator A and closed curve Γ as described in the proof of Theorem 4.4.6. Then by [51, Chapter IV, Theorem 3.1 & Remark 3.3], as $\Gamma \subset \rho(A)$ is compact, there exists some $\varepsilon_0 > 0$ such that for any $A_\varepsilon \in \mathscr{B}(\mathcal{H})$ with $\|A - A_\varepsilon\|_{\mathrm{op}} < \varepsilon_0$, the curve Γ is contained in the resolvent $\rho(A_\varepsilon)$ and contains $\sigma(A_\varepsilon)$ in its enclosure. Then by applying Theorem 4.4.6, the result follows. □

So, a family of operators $(A_\varepsilon) \subset \mathscr{B}(\mathcal{H})$ that converges in operator norm to some $A \in \mathscr{B}(\mathcal{H})$ as ε vanishes, with the properties as described in Theorem 4.4.6, has that the Krylov solvability of $A_\varepsilon f_\varepsilon = g$ survives in the vanishing limit.

Although the problem $Af_\varepsilon = g_\varepsilon$, for a perturbed datum g_ε, may be constructed in the form $A_\varepsilon f_\varepsilon = g$, in general $\overline{\mathcal{K}(A, g_\varepsilon)} \neq \overline{\mathcal{K}(A_\varepsilon, g)}$. So *exactly* the Krylov space to consider is also a non-trivial question.

To unmask these ideas a few small, yet informative, examples are presented in what follows.

Example 7.2.2. This first example considers the situation of Krylov solvability being *lost* in the limit of $\varepsilon \to 0$.

Consider the unweighted right-shift operator $\mathcal{R}$ on $\ell^2(\mathbb{Z})$ (see Appendix A). This operator is unitary and cyclic. It is also known that for the vector $g = e_1$, the solution to $\mathcal{R}f = g$, $f = e_0$, is perpendicular to $\overline{\mathcal{K}(\mathcal{R}, g)}$. Yet, as $\mathcal{R}$ is cyclic, the set of cyclic vectors is *dense* in $\mathcal{H}$ [40]. So, choosing a vector $g_\varepsilon \in \mathcal{H}$ that is cyclic and arbitrarily close to g in the $\mathcal{H}$-norm generates the whole Hilbert space. The problem under this perturbation is

$$\mathcal{R}f_\varepsilon = g_\varepsilon$$

which is obviously Krylov solvable. Moreover, this may actually be re-cast in terms of a perturbation of the operator $\mathcal{R}$ as was stated above. Indeed, let $v = g_\varepsilon - g$, so that

$$\left(\mathcal{R} - \frac{1}{\|f_\varepsilon\|^2_{\mathcal{H}}} |v\rangle \langle f_\varepsilon|\right) f_\varepsilon = \mathcal{R}_\varepsilon f_\varepsilon = g\,.$$

But considering the Krylov spaces generated from perturbation of the datum, $f_\varepsilon \in \overline{\mathcal{K}(\mathcal{R}, g_\varepsilon)}$, and yet $f \notin \overline{\mathcal{K}(\mathcal{R}, g)}$ when $\varepsilon = 0$. This is a case of when the ε-Krylov solvable problem is not Krylov solvable in the limit $\varepsilon \to 0$.

Example 7.2.3. Let $A \in \mathscr{B}(\mathcal{H})$ be self-adjoint, positive definite and such that $0 \in \sigma(A)$, with $\sigma(A) \subset [0, \|A\|_{\mathrm{op}}]$. Consider the linear inverse problem $Af = g$ with $g \in \mathrm{ran}A$, and some $\varepsilon > 0$.

Let $A_\varepsilon = A + \varepsilon\mathbb{1}$, so that $0 \in \rho(A_\varepsilon)$. Then consider the perturbed linear inverse problem $A_\varepsilon f_\varepsilon = g$. This is a case of both the perturbed problem and the unperturbed problem being Krylov solvable, i.e., one has $f_\varepsilon \in \overline{\mathcal{K}(A_\varepsilon, g)}$ *and* $f \in \overline{\mathcal{K}(A, g)}$. So, in this case, the ε-Krylov solvable problem remains Krylov solvable in the limit $\varepsilon \to 0$.

Example 7.2.4. For concreteness, consider the operator $A : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ with the properties as described in Example 7.2.3, defined by $e_n \mapsto \frac{1}{n}e_n$

for all $n \in \mathbb{N}$, where $(e_n)_{n\in\mathbb{N}}$ is the canonical basis. Take some generic $g \in \mathrm{ran}A$, namely $g = \sum_{n\in\mathbb{N}} g_n e_n$, so that clearly for the linear inverse problem $Af = g$, one has the solution $f = \sum_{n\in\mathbb{N}} n g_n e_n$. As $g \in \mathrm{ran}A$, clearly $\sum_{n\in\mathbb{N}} n^2|g_n|^2 < \infty$.

Now considering the perturbed problem $A_\varepsilon f_\varepsilon = g$, where $A_\varepsilon = A + \varepsilon\mathbb{1}$, the solution to the perturbed case is $f_\varepsilon = \sum_{n\in\mathbb{N}}(\varepsilon + 1/n)^{-1} g_n e_n$. Comparing the solution to the perturbed an unperturbed problem, one has that $\|f - f_\varepsilon\|_{\mathcal{H}} \to 0$ as $\varepsilon \to 0$. Indeed,

$$f - f_\varepsilon = \sum_{n\in\mathbb{N}} \left(n - \frac{1}{\varepsilon + \frac{1}{n}} \right) g_n e_n \,,$$

and $n - (\varepsilon + 1/n)^{-1} = \varepsilon n^2/(1 + \varepsilon n)$ so that

$$\|f - f_\varepsilon\|^2_{\mathcal{H}} = \sum_{n\in\mathbb{N}} \frac{\varepsilon^2 n^4}{(1+\varepsilon n)^2} |g_n|^2 \,.$$

As $n^4/(1+\varepsilon n)^2 = n^2/(1/\varepsilon + n)^2 \leq n^2$, due to dominated convergence, one has that $\|f - f_\varepsilon\|^2_{\mathcal{H}} \to 0$ as $\varepsilon \to 0$.

The above examples illustrate two scenarios:

(i) ε-dependent problems that are Krylov solvable for $\varepsilon > 0$ *and* when $\varepsilon = 0$.

(ii) ε-dependent problems that are Krylov solvable for $\varepsilon > 0$ but *not* Krylov solvable for $\varepsilon = 0$.

Example 7.2.4 also unmasks the situation where there is *strong* convergence of the perturbed solution f_ε to the unperturbed solution f, although the unperturbed operator A has an unbounded inverse.

This theory at such a general level presents an interesting area to explore, and develop conditions by which one may investigate the Krylov solvability of linear inverse problems by an auxiliary (i.e., perturbed) equation. An advantage being that the perturbed equation may be simpler to analyse than the unperturbed problem.

Some particular applications of the development of this theory could include the investigation of what are known as boundary layer problems using Krylov methods. The differential systems that give rise to these problems are classically known as *singularly perturbed equations.* This name arises from the fact that the solution to the perturbed problem in the vanishing limit $\varepsilon \to 0$ does not approach the unperturbed solution $\varepsilon = 0$ in an appropriate norm [95]. Under these singular perturbations of an operator A to A_ε, the nature of the equation $Af = g$ fundamentally changes, hence the appearance of a boundary layer where the derivative changes rapidly [8, 95]. Some classical examples from physical systems include flow around an aerofoil [60] and also thermal boundary layer problems present in heat transfer [47]. An example of this in quantum physics is the perturbation of the Schrödinger equation by the Laplacian, known in this context as a "semi-classical" limit [62].

7.3 Krylov methods applied to Friedrichs systems

Another area that is planned for an investigation using Krylov methods is that of the Krylov solvability of Friedrich systems. Friedrichs systems are a class of densely defined linear differential operators. In the *abstract* complex Hilbert space setting, a Friedrichs system is formulated as follows. Let $\mathcal{D} \subset \mathcal{H}$ be a dense subspace and $T, \widetilde{T} : \mathcal{D} \to \mathcal{H}$ be linear operators satisfying

$$\langle \psi, T\varphi \rangle = \left\langle \widetilde{T}\psi, \varphi \right\rangle \quad \forall\, \varphi, \psi \in \mathcal{D}\,, \tag{7.2}$$

$$\exists\, c > 0 \text{ such that } \forall\, \varphi \in \mathcal{D} \quad \left\| \left(T + \widetilde{T}\right) \varphi \right\|_{\mathcal{H}} \leq c\, \|\varphi\|_{\mathcal{H}}\,, \tag{7.3}$$

usually in addition to a coercivity-type condition

$$\left\langle \varphi, \left(T + \widetilde{T}\right) \varphi \right\rangle \geq 2\mu_0\, \|\varphi\|^2_{\mathcal{H}} \quad \forall\, \varphi \in \mathcal{D} \tag{7.4}$$

for some $\mu_0 > 0$ [3, 2]. Some nice examples of Friedrichs systems may be found in [3]. As typically Friedrichs systems are differential operators, the operators T, $\widetilde{T}$ defined by the conditions (7.2) and (7.3) are assumed to be unbounded.

Certainly, conditions have been developed to permit the treatment of these types of equations with finite-element methods under particular boundary conditions [29]. These conditions amount to the restriction of the operator to certain subspaces that ensure one has a homeomorphism, which is then suitable for treatment with finite-element methods. But unbounded Krylov methods may be able to provide new insights and methods into equations that are unsuitable for treatment with finite-element techniques. This would effectively mean that one could remove conditions on the boundary that guarantee boundedness and invertability of the operator.

Appendix A

Frequently Used Operators

Multiplication operator on $\ell^2(\mathbb{N})$

Let $(e_n)_{n\in\mathbb{N}}$ denote the canonical orthonormal basis of $\ell^2(\mathbb{N})$. For a given bounded sequence $a \equiv (a_n)_{n\in\mathbb{N}} \subset \mathbb{C}$, the multiplication operator is $M^{(a)} : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ and its action defined by $M^{(a)}e_n = a_n e_n \, \forall n \in \mathbb{N}$, then extended by linearity and density of the basis in $\ell^2(\mathbb{N})$. Hence

$$M^{(a)} = \sum_{n=1}^{\infty} a_n \left| e_n \right\rangle \left\langle e_n \right| \, , \tag{A.1}$$

that converges in the strong operator topology. $M^{(a)}$ is bounded with norm $\left\| M^{(a)} \right\|_{\mathrm{op}} = \sup_n |a_n|$ and the spectrum $\sigma(M) = \overline{\{a_n\}_{n\in\mathbb{N}}}$. The adjoint is the multiplication operator defined by the conjugate sequence $a^* \equiv (a_n^*)_{n\in\mathbb{N}}$, i.e., $M^* = M^{(a^*)}$, and thus $M^{(a)}$ is a normal operator. If $a = a^*$, then one immediately has that $M^{(a)}$ is self-adjoint, and if $\lim_{n\to\infty} a_n = 0$, then $M^{(a)}$ is compact.

Right- & left-shift operator on $\ell^2(\mathbb{N})$

The *right*-shift operator $R : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ is defined by $Re_n = e_{n+1}$ for all $n \in \mathbb{N}$, then extended by linearity and density to $\ell^2(\mathbb{N})$, i.e.,

$$R = \sum_{n=1}^{\infty} \left| e_{n+1} \right\rangle \left\langle e_n \right| \, , \tag{A.2}$$

that converges in the strong operator topology. R is an isometry, i.e., $\|Ru\|_{\ell^2(\mathbb{N})} = \|u\|_{\ell^2(\mathbb{N})}$ for all $u \in \ell^2(\mathbb{N})$, and $\mathrm{ran}R = \{e_1\}^\perp$. It is bounded with $\|R\|_{\mathrm{op}} = 1$, injective, and invertible on its range with bounded inverse

$$R^{-1} : \mathrm{ran}R \to \ell^2(\mathbb{N}) \,, \quad R^{-1} = \sum_{n=1}^{\infty} |e_n\rangle \langle e_{n+1}| \;. \tag{A.3}$$

The adjoint of R on $\ell^2(\mathbb{N})$ is the left-shift operator $L : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$, defined by $Le_{n+1} = e_n$, again extended by linearity and density to $\ell^2(\mathbb{N})$, i.e.,

$$L = \sum_{n=1}^{\infty} |e_n\rangle \langle e_{n+1}| \;, \quad R^* = L \,. \tag{A.4}$$

L inverts R on its range, i.e., $LR = \mathbb{1}$, but $RL = \mathbb{1} - |e_1\rangle \langle e_1|$. Hence $\ker L = \ker R^* = \mathrm{span}\,\{e_1\}$. Also, $\|L\|_{\mathrm{op}} = 1$.

Both L and R have the same spectrum $\sigma(L) = \sigma(R) = \{z \in \mathbb{C};\, |z| \le 1\}$. R has no eigenvalues, whereas the eigenvalues of L form the open unit ball $\{z \in \mathbb{C};\, |z| < 1\}$.

Clearly the resolvent sets for L and R contain the set $\{z \in \mathbb{C};\, |z| > 1\}$. This is a consequence of $\|R\|_{\mathrm{op}} = \|L\|_{\mathrm{op}} = 1$.

To show that every value of $z \in \mathbb{C}$ such that $|z| < 1$ is an eigenvalue of L, consider the vector $u \in \ell^2(\mathbb{N})$

$$u = \sum_{n \in \mathbb{N}} z^n e_n \,,$$

then one has $Lu = z \sum_{n \in \mathbb{N}} z^n e_n = zu$, so that $z \in \sigma(L)$. As the spectrum is closed, then clearly $\sigma(L)$ is the closed unit disc.

To show that the right-shift has no eigenvalues for $|z| \le 1$, one may consider the generic vector $u \in \ell^2(\mathbb{N})$

$$u = \sum_{n \in \mathbb{N}} u_n e_n \,,$$

and by contradiction, assume that there is some $\ell^2(\mathbb{N})$ sequence $(u_n)_{n \in \mathbb{N}}$ such that $Ru = zu$. One obtains the recurrence relation for the sequence $(u_n)_{n\in\mathbb{N}}$

$$u_{n+1} = zu_n \,,$$

so that $u_n = z^{-1}u_{n+1}$. Then

$$u = \sum_{n\in\mathbb{N}} z^{-n+1}u_1 e_1 \,,$$

which is not in $\ell^2(\mathbb{N})$ (unless $u = 0$) as $z^{-n+1} \to \infty$ as $n \to \infty$. Therefore R has no eigenvalues.

Showing that the spectrum of R is the unit disc requires more work. Consider any $u \in \ell^2(\mathbb{N})$ and $|z| < 1$, with

$$u = \sum_{n\in\mathbb{N}} u_n e_n \,,$$

and by contradiction, assume that $R - z\mathbb{1}$ is a bijection, i.e., $z \in \rho(R)$. Choose any $v \in \ell^2(\mathbb{N})$,

$$v = \sum_{n\in\mathbb{N}} v_n e_n \,,$$

and consider $(R - z\mathbb{1})u = v$, which has a unique solution $u \in \ell^2(\mathbb{N})$. Immediately one has the following

$$(R - z\mathbb{1})u = \sum_{n\geq 2}(u_{n-1} - zu_n)e_n - zu_1e_1 = v \,,$$

that leads to the following recurrence formula $v_1 = -zu_1$, $v_2 = u_1 - zu_2$, etc, or more generally

$$v_n = -\sum_{i=1}^{n-1} \frac{1}{z^{n-i}} v_i - zu_n \,.$$

Clearly, $u_n \to 0$ as $n \to \infty$. As $v \in \ell^2(\mathbb{N})$ is arbitrary, choose $v_i = z^i$ that ensures $v \in \ell^2(\mathbb{N})$. In this particular case the following relationship is obtained for u_n.

$$z^n = v_n = \sum_{i=1}^{n-1} \frac{1}{z^n} - zu_n \,,$$

however, as $\frac{1}{|z^n|} \to \infty$ as $n \to \infty$, it is clear that $u_n \to \infty$ as $n \to \infty$ which violates $u \in \ell^2(\mathbb{N})$. Therefore, $R - z\mathbb{1}$ is *not* bijective for $|z| < 1$. As the spectrum is closed, $\sigma(R)$ is the closed unit disc in $\mathbb{C}$.

Compact (weighted) left- & right-shift operator on $\ell^2(\mathbb{N})$

This is the operator $\mathcal{R} : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ defined by

$$\mathcal{R} = \sum_{n=1}^{\infty} \sigma_n \left|e_{n+1}\right\rangle \left\langle e_n\right| \,, \tag{A.5}$$

that converges in the operator norm, where $\sigma \equiv (\sigma_n)_{n\in\mathbb{N}}$ is a given bounded sequence such that $0 < \sigma_{n+1} < \sigma_n$ $\forall n \in \mathbb{N}$ and $\lim_{n\to\infty} \sigma_n = 0$. $\mathcal{R}$ is injective and compact with (A.5) the singular value decomposition, and $\|\mathcal{R}\|_{\text{op}} = \sigma_1$, ran$\mathcal{R} = \{e_1\}^{\perp}$. The adjoint of $\mathcal{R}$ is the compact left-shift operator $\mathcal{L} : \ell^2(\mathbb{N}) \to \ell^2(\mathbb{N})$ defined by

$$\mathcal{L} = \sum_{n=1}^{\infty} \sigma_n \left|e_n\right\rangle \left\langle e_{n+1}\right| = \mathcal{R}^* \,, \tag{A.6}$$

where convergence of the series occurs in the operator norm. Moreover, $\mathcal{L}\mathcal{R} = M^{(\sigma^2)}$ and $\mathcal{R}\mathcal{L} = M^{(\sigma^2)} - \sigma_1^2 \left|e_1\right\rangle \left\langle e_1\right|$. Also $\|\mathcal{L}\|_{\text{op}} = \sigma_1$.

Weighted (compact) left- & right-shift operator on $\ell^2(\mathbb{Z})$

The compact right shift is the operator $\mathcal{R} : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$ defined by the operator norm convergent series

$$\mathcal{R} = \sum_{n\in\mathbb{Z}} \sigma_{|n|} \left|e_{n+1}\right\rangle \left\langle e_n\right| \,, \tag{A.7}$$

where $\sigma \equiv (\sigma_n)_{n\in\mathbb{N}}$ is a given bounded sequence such that $0 < \sigma_{n+1} < \sigma_n$ $\forall n \in \mathbb{N}$ and $\lim_{n\to\infty} \sigma_n = 0$. $\mathcal{R}$ is injective and compact with ran$\mathcal{R}$ dense in $\ell^2(\mathbb{Z})$ and norm $\|\mathcal{R}\|_{\text{op}} = \sigma_0$. (A.7) is the singular value decomposition of $\mathcal{R}$.

The adjoint of $\mathcal{R}$ is the compact left shift on $\ell^2(\mathbb{Z})$

$$\mathcal{R}^* = \mathcal{L} = \sum_{n\in\mathbb{Z}} \sigma_{|n|} \left|e_n\right\rangle \left\langle e_{n+1}\right| \, . \tag{A.8}$$

Also, $\mathcal{L}\mathcal{R} = M^{(\sigma^2)} = \mathcal{R}\mathcal{L}$, and $\mathcal{L}$ has dense range.

The inverse operators of both $\mathcal{R}$ and $\mathcal{L}$ are densely defined, surjective, unbounded operators with actions given by the strong operator topology convergent series

$$\mathcal{R}^{-1} = \sum_{n\in\mathbb{Z}} \frac{1}{\sigma_{|n|}} \left|e_n\right\rangle \left\langle e_{n+1}\right| \, , \quad \mathcal{L}^{-1} = \sum_{n\in\mathbb{Z}} \frac{1}{\sigma_{|n|}} \left|e_{n+1}\right\rangle \left\langle e_n\right| \, . \tag{A.9}$$

Volterra operator on $L^2[0,1]$

The Volterra operator $V : L^2[0,1] \to L^2[0,1]$ is defined by the integral

$$(Vf)(x) = \int_0^x f(y)\,\mathrm{d}y \quad x \in [0,1] \, . \tag{A.10}$$

V is a compact, injective operator with spectrum $\sigma(V) = \{0\}$, and norm $\|V\|_{\text{op}} = \frac{2}{\pi}$. The adjoint V^* acts as

$$(V^*f)(x) = \int_x^1 f(y)\,\mathrm{d}y \, , \quad x \in [0,1] \, , \tag{A.11}$$

so that $V + V^*$ is the rank-one orthogonal projection

$$V + V^* = \left|\mathbf{1}\right\rangle \left\langle \mathbf{1}\right| \, , \tag{A.12}$$

onto the function $\mathbf{1}(x) = 1$. The singular value decomposition of V is

$$V = \sum_{n=0}^{\infty} \sigma_n \left|\psi_n\right\rangle \left\langle \varphi_n\right| \, , \qquad \begin{array}{rcl} \sigma_n &=& \frac{2}{(2n+1)\pi} \\ \varphi_n(x) &=& \sqrt{2}\,\cos\frac{(2n+1)\pi}{2}x \\ \psi_n(x) &=& \sqrt{2}\,\sin\frac{(2n+1)\pi}{2}x \, , \end{array} \tag{A.13}$$

where both $(\psi_n)_{n\in\mathbb{N}_0}$ and $(\varphi_n)_{n\in\mathbb{N}_0}$ are orthonormal bases for $L^2[0,1]$. Therefore, ran$V$ is dense, but strictly contained in $L^2[0,1]$, for example $\mathbf{1} \notin$ ran$V$.

The resolvent function $(z\mathbb{1} - V)^{-1}$ for $z \in \mathbb{C} \setminus \{0\}$ is expressed by

$$(z\mathbb{1} - V)^{-1}\psi = z^{-1}\psi + z^{-2} \int_0^x e^{\frac{x-y}{z}} \psi(y)\,\mathrm{d}y\,, \tag{A.14}$$

for all $\psi \in L^2[0,1]$. The explicit action of powers of V is given by

$$(V^n f)(x) = \frac{1}{(n-1)!} \int_0^x (x-y)^{n-1} f(y)\,\mathrm{d}y\,, \quad n \in \mathbb{N}\,. \tag{A.15}$$

Multiplication operator on a disc in $L^2(\Omega)$

The multiplication operator $M_z : L^2(\Omega) \to L^2(\Omega)$ is defined by the action $f \mapsto zf$, where

$$\Omega := \left\{ z \in \mathbb{C};\ \left| z - \frac{3}{4} \right| < \frac{1}{2} \right\}\,. \tag{A.16}$$

M_z is normal, with adjoint $(M_z)^* = M_{\overline{z}}$, norm $\|M_z\|_{\mathrm{op}} = 1$, and spectrum $\sigma(M_z) = \overline{\Omega}$ with no eigenvalues. Furthermore, M_z is a bijection, as $\frac{1}{z} \in L^2(\Omega)$ and $M_z^{-1} = M_{\frac{1}{z}}$.

Firstly, consider $\zeta \in \mathbb{C} \setminus \overline{\Omega}$. Then given any $f \in L^2(\Omega)$, one has that $(\zeta\mathbb{1} - M_z)f = (\zeta - z)f \in L^2(\Omega)$. As such, $(\zeta\mathbb{1} - M_z)^{-1} = M_{(\zeta-z)^{-1}}$ and this is a bounded linear operator, meaning that $\zeta \in \rho(M_z)$. One may see this by the fact that $|\zeta - z| \geq \operatorname{dist}(\zeta, \overline{\Omega}) > 0$ ensuring that $|\zeta - z|^{-1} \leq \operatorname{dist}(\zeta, \overline{\Omega})^{-1} < \infty$, and therefore given any $g \in L^2(\Omega)$, one has that $(\zeta - z)^{-1} g \in L^2(\Omega)$.

If, on the other hand, $\zeta \in \overline{\Omega}$, then taking $g \in L^2(\Omega)$ to be $g = \sqrt{\zeta - z}$, then $(\zeta\mathbb{1} - M_z)^{-1} g = \frac{1}{\sqrt{\zeta - z}} \notin L^2(\Omega)$. Therefore, $\zeta \in \sigma(M_z)$.

Lastly, to see that there are indeed no eigenvalues of M_z, by contradiction assume that $f \neq 0$ a.e. is an eigenvector with eigenvalue λ. Then $M_z f = \lambda f$ which implies that $(\lambda - z)f = 0$, and therefore $z = \lambda$ on the support of f in Ω, which is *not* of zero measure. This is impossible, so M_z has no eigenvalues.

Appendix B

Operator Theory Miscellanea

The definitions, theorems and proofs here are intended to provide an outline to some operator theoretic notions to supplement the materials presented in Chapters 4, 5, and 6. The basic outlines, along with some proofs of fundamental results, are illsutrated, of course with no pretention of providing a full account of the topics discussed herein. For a full account, the reader is referred to the following monographs [88, 51, 77, 6, 76].

B.1 The graph of an operator

Definition B.1.1. The graph space of a linear operator $A : \mathcal{H} \to \mathcal{H}$ with domain $\mathcal{D}(A)$ in Hilbert space $\mathcal{H}$ is the subset

$$G(A) := \{(x, Ax) \,|\, x \in \mathcal{D}(A)\} \tag{B.1}$$

of the Banach space $\mathcal{H} \times \mathcal{H}$, where $\mathcal{H} \times \mathcal{H}$ has the metric $\|(x, y)\|^2_{\mathcal{H}\times\mathcal{H}} = \|x\|^2_{\mathcal{H}} + \|y\|^2_{\mathcal{H}}$. The corresponding *graph norm* defined on $G(A)$ is $\|\cdot\|^2_{G(A)} := \|\cdot\|^2_{\mathcal{H}} + \|A\cdot\|^2_{\mathcal{H}}$. An operator A is said to be *closed* if $G(A) = \overline{G(A)}$.

The graph space inherits the subspace topology of $\mathcal{H} \times \mathcal{H}$. Many standard properties of closed operators and graph spaces may be found in standard functional analysis texts (e.g., [10, 51]).

Remark B.1.2. For any linear operator $A : \mathcal{H} \to \mathcal{H}$, its closedness it tantamount to saying that for every convergent sequence $((u_n, Au_n))_{n\in\mathbb{N}} \subset G(A)$ one has

$$\begin{cases} u_n \to u \in \mathcal{D}(A) \\ Au_n \to v \in \mathcal{H} \end{cases} , \quad \text{where } v = Au\,. \tag{B.2}$$

This is equivalent to saying that the graph space of a closed operator is a closed linear submanifold of $\mathcal{H} \times \mathcal{H}$.

Definition B.1.3. A linear operator $A : \mathcal{H} \to \mathcal{H}$ with domain $\mathcal{D}(A)$ in Hilbert space $\mathcal{H}$ is said to be *closable* if there exists an extension of A, that shall be denoted by $\overline{A}$, the closure of A, such that $\overline{G(A)} = G\left(\overline{A}\right)$.

Remark B.1.4. An operator A is closable if and only if there is no element $v \in \mathcal{H} \setminus \{0\}$ such that $(0, v) \in G\left(\overline{A}\right)$ [51, Chapter III, Section 5.3]. Obviously, if A is closable, then $A \subset \overline{A}$.

B.2 The spectral integral for self-adjoint operators

The main aim of this Section is to provide a reasonably self-contained background for the development of spectral integrals of self-adjoint operators and the functional calculus. Further details along with the proofs may be found in [88, Chapters 4 & 5].

To begin with, the well-known *polarisation* formula is presented.

Lemma B.2.1. *If $A : \mathcal{H} \to \mathcal{H}$ is a linear operator on Hilbert space $\mathcal{H}$, then*

$$\begin{aligned} 4\langle y,\, Ax\rangle = &\langle x+y,\, A(x+y)\rangle - \langle x-y,\, A(x-y)\rangle \\ &+ i\langle x+iy,\, A(x+iy)\rangle - i\langle x-iy,\, A(x-iy)\rangle\ , \end{aligned} \tag{B.3}$$

for all $x, y \in \mathcal{D}(A)$.

Proof. See [88, equation (1.2)] or [51, Chapter I, Problem 6.13]. □

B.2.1 The spectral measure

Definition B.2.2 (Definition 4.1 [88], [51])**.** A family of orthogonal projections $\{\mathbf{E}(\lambda) \mid \lambda \in \mathbb{R}\}$ on the Hilbert space $\mathcal{H}$ is called a *resolution of the identity* if

(i) $\mathbf{E}(\lambda)\mathbf{E}(\lambda') = \mathbf{E}(\lambda')\mathbf{E}(\lambda) = \mathbf{E}(\min\{\lambda, \lambda'\})$

(ii) $\lim_{\lambda\to\infty} \mathbf{E}(\lambda) u = u$ and $\lim_{\lambda\to-\infty} \mathbf{E}(\lambda) u = 0$ for all $u \in \mathcal{H}$.

Definition B.2.3 (Definition 4.2 [88])**.** Let $\mathfrak{M}$ be an algebra of subsets of Ω, and $\mathcal{H}$ a Hilbert space. A spectral premeasure on $\mathfrak{M}$ is a mapping $\mathbf{E}$ of $\mathfrak{M}$ into the orthogonal projections on $\mathcal{H}$ such that

(i) $\mathbf{E}(\Omega) = \mathbb{1}$,

(ii) $\mathbf{E}$ is strongly countably additive.

If $\mathfrak{M}$ is a σ-algebra, then $\mathbf{E}$ is a *spectral measure* on $\mathfrak{M}$. Note: Infinite sums $\sum_{n\in\mathbb{N}} \mathbf{E}(M_n)$ for $M_n \in \mathfrak{M}$ are meant in the sense of strong operator topology convergence, e.g. for $(M_n)_{n\in\mathbb{N}}$ a disjoint collection of sets with $M_n \in \mathfrak{M}$ for all $n \in \mathbb{N}$, one has $\mathbf{E}\left(\bigcup_{n\in\mathbb{N}} M_n\right) u = \lim_{n\to\infty} \sum_{n=1}^{N} \mathbf{E}(M_n) u$ for $u \in \mathcal{H}$.

The property (ii) above clearly shows that $\mathbf{E}(\emptyset) = \mathbb{O}$, and that $\mathbf{E}$ is finitely additive. Moreover, the spectral premeasure can be extended to a spectral measure in the following lemma.

Lemma B.2.4 (Lemma 4.9 [88])**.** *Let $\mathbf{E}_0$ be a spectral premeasure on an algebra $\mathfrak{M}_0$ of subsets of a set Ω. Then there is a spectral measure $\mathbf{E}$ on the σ-algebra $\mathfrak{M}$ generated by $\mathfrak{M}_0$ such that $\mathbf{E}_0(M) = \mathbf{E}(M)$ for all $M \in \mathfrak{M}_0$.*

Lemma B.2.5 (Lemma 4.3 [88])**.** *If $\mathbf{E}$ is a finitely additive map of an algebra $\mathfrak{M}$ into the orthogonal projections onto Hilbert space, then for $M, N \in \mathfrak{M}$*

$$\mathbf{E}(M)\mathbf{E}(N) = \mathbf{E}(M \cap N)\ . \tag{B.4}$$

The next lemma, as presented in [88], shows that the spectral measure may be used to construct a *scalar* measure on the σ-algebra $\mathfrak{M}$.

Lemma B.2.6 (Lemma 4.4 [88])**.** *A map* $\mathbf{E}$ *of a* σ*-algebra* $\mathfrak{M}$ *on a set* Ω *into the orthogonal projections on* $\mathcal{H}$ *is a spectral measure if and only if* $\mathbf{E}(\Omega) = \mathbb{1}$ *and for each vector* $x \in \mathcal{H}$*, the set function* $\mu_x(\cdot) := \langle x, \mathbf{E}(\cdot)\, x\rangle$ *on* $\mathfrak{M}$ *is a measure.*

Proof. The 'only if' part is immediate from Definition B.2.3. For the 'if' part, consider $(M_n)_{n\in\mathbb{N}}$ a sequence of disjoint sets in $\mathfrak{M}$, and let $M = \bigcup_{n\in\mathbb{N}} M_n$. $\mu_x(\cdot)$ is countably additive because by assumption it is a measure. So

$$\begin{aligned}\mu_x(M) = \mu_x\left(\bigcup_{n\in\mathbb{N}} M_n\right) = \left\langle x, \mathbf{E}\left(\bigcup_{n\in\mathbb{N}} M_n\right) x\right\rangle &= \sum_{n\in\mathbb{N}} \mu_x(M_n)\\ &= \sum_{n\in\mathbb{N}} \langle x, \mathbf{E}(M_n)\, x\rangle\ ,\end{aligned}$$

so that for every $x \in \mathcal{H}$,

$$\left\langle x, \mathbf{E}\left(\bigcup_{n\in\mathbb{N}} M_n\right) x\right\rangle = \sum_{n\in\mathbb{N}} \langle x, \mathbf{E}(M_n)\, x\rangle\ .$$

By the polarisation formula (B.3) one has that, in the strong operator topology,

$$\mathbf{E}(M) = \sum_{n\in\mathbb{N}} \mathbf{E}(M_n)\ .$$

From the definition of a spectral measure, the proof is complete. □

The next theorem shows the crucial connection between the spectral integral and the resolution of the identity.

Theorem B.2.7 (Theorem 4.6 [88])**.** *If* $\mathbf{E}$ *is a spectral measure on the Borel* σ*-algebra* $\mathfrak{B}(\mathbb{R})$ *in* $\mathcal{H}$*, then for* $\lambda \in \mathbb{R}$

$$\mathbf{E}(\lambda) := \mathbf{E}((-\infty, \lambda]) \tag{B.5}$$

defines a resolution of the identity. Conversely, for each resolution of the identity, there is a unique spectral measure $\mathbf{E}$ *on* $\mathfrak{B}(\mathbb{R})$ *such that* (B.5) *is true.*

The following lemmas pave the way for the next section on the spectral integral.

Lemma B.2.8 (Lemma 4.8 [88])**.** *Let* $\mathbf{E}$ *be a spectral measure on* $(\Omega, \mathfrak{M})$ *in a Hilbert space* $\mathcal{H}$*. Define* $\mu_{x,y}(M) := \langle y, \, \mathbf{E}\,(M)\,x\rangle$ *and* $\mu_x\,(M) := \langle x, \, \mathbf{E}\,(M)\,x\rangle$ *for* $M \in \mathfrak{M}$*. Then*

(i) $|\mu_{x,y}|(M) \leq \sqrt{\mu_x\,(M)\,\mu_y\,(M)}$ *for* $x, y \in \mathcal{H}$ *and* $M \in \mathfrak{M}$*.*

(ii) If $h \in L^2(\Omega, \mu_x)$ *and* $g \in L^2(\Omega, \mu_y)$*, then*

$$\left|\int_\Omega hg \, \mathrm{d}\mu_{x,y}\right| \leq \int_\Omega |hg| \, \mathrm{d}|\mu_{x,y}| \leq \|h\|_{L^2(\Omega,\mu_x)} \, \|g\|_{L^2(\Omega,\mu_y)} \, . \tag{B.6}$$

Proof. For part (i), let $M \in \mathfrak{M}$ be the disjoint union of a countable collection of sets $(M_n)_{n\in\mathbb{N}}$. Then from Cauchy-Schwartz

$$\begin{aligned} |\mu_{x,y}\,(M_n)\,| &= |\langle \mathbf{E}\,(M_n)\,y, \, \mathbf{E}\,(M_n)\,x\rangle| \\ &\leq \|\mathbf{E}\,(M_n)\,x\|_{\mathcal{H}} \, \|\mathbf{E}\,(M_n)\,y\|_{\mathcal{H}} = \mu_x\,(M_n)^{\frac{1}{2}} \, \mu_y\,(M_n)^{\frac{1}{2}} \ , \end{aligned}$$

and using Cauchy-Schwartz again, along with the countable additivity of the measure, one has

$$\sum_{n\in\mathbb{N}} |\mu_{x,y}\,(M_n)\,| \leq \sqrt{\mu_x\,(M)\,\mu_y\,(M)} \, ,$$

and taking the supremum over the disjoint partitions $(M_n)_{n\in\mathbb{N}}$ that give M, one finally has $|\mu_{x,y}|(M) \leq \sqrt{\mu_x\,(M)\,\mu_y\,(M)}$.

For part (ii) only the simple functions are considered as they are dense in L^2 spaces. Taking f and g as simple functions, using part (i) and Cauchy-Schwartz, the result immediately follows. □

Definition B.2.9 (Definition 4.3 [88])**.** The support of a spectral measure $\mathbf{E}$ on the Borel σ-algebra $\mathfrak{B}(\Omega)$ is the complement in Ω of the union of all open sets $\mathcal{U} \subset \Omega$ such that $\mathbf{E}\,(\mathcal{U}) = \mathbb{O}$.

B.2.2 The spectral integral

From the construction of a spectral measure, it is immediate that one may define an integral. Much of the following theory regarding the spectral integrals of certain functions is built by using simple approximating function, exactly as done in the construction of the Lebesgue integral.

Definition B.2.10. Let $\mathfrak{M}$ be a σ-algebra of subsets of a set Ω, and $\mathbf{E}$ a spectral measure on $(\Omega, \mathfrak{M})$. Then the operator known as the *spectral integral* of an $\mathbf{E}$ almost everywhere finite, $\mathfrak{M}$-measurable function $f : \Omega \to \mathbb{C} \cup \{\infty\}$ is defined as follows.

$$\mathbb{I}\left(f\right) := \int_{\Omega} f(t)\,\mathrm{d}\mathbf{E}\left(t\right) \ . \tag{B.7}$$

The space $\mathcal{S} = \mathcal{S}(\Omega, \mathfrak{M}, \mathbf{E})$ is the space of $\mathbf{E}$ almost everywhere ($\mathbf{E}$-a.e.) finite, $\mathfrak{M}$-measurable functions f on Ω.

Note that the $\mathbf{E}$ almost everywhere finiteness of a function f means that $\mathbf{E}\left(\{t \in \Omega \,|\, f(t) = \infty\}\right) = \mathbb{O}$.

Next, the concept of a bounding sequence is introduced for the purposes of some of the subsequent results.

Definition B.2.11 (Definition 4.4 [88])**.** A sequence of sets $(M_n)_{n\in\mathbb{N}} \subset \mathfrak{M}$, of a σ-algebra $\mathfrak{M}$, is a bounding sequence for some subset of functions $\mathcal{F} \subset \mathcal{S}(\Omega, \mathfrak{M}, \mathbf{E})$ if each function $f \in \mathcal{F}$ is bounded on M_n and $M_n \subset M_{n+1}$ for all $n \in \mathbb{N}$, and $\mathbf{E}\left(\bigcup_{n\in\mathbb{N}} M_n\right) = \mathbb{1}$.

For any set of finite elements of $\mathcal{S}(\Omega, \mathfrak{M}, \mathbf{E})$ one may show a bounding sequence exists (see comments after Definition 4.4 [88]).

Theorem B.2.12 (Theorem 4.13 [88])**.** *Let $(M_n)_{n\in\mathbb{N}}$ be a bounding sequence for a function $f \in \mathcal{S}(\Omega, \mathfrak{M}, \mathbf{E})$. Then one has:*

(i) A vector $v \in \mathcal{H}$ is in $\mathcal{D}\left(\mathbb{I}\left(f\right)\right)$ if and only if the sequence $\left(\mathbb{I}\left(f\chi_{M_n}\right)v\right)_{n\in\mathbb{N}}$ converges in $\mathcal{H}$, or equivalently, if $\sup_{n\in\mathbb{N}} \left\|\mathbb{I}\left(f\chi_{M_n}\right)v\right\|_{\mathcal{H}} < \infty$.

(ii) For $v \in \mathcal{D}(\mathbb{I}(f))$, the limit sequence $(\mathbb{I}(f)v)_{n\in\mathbb{N}}$ does not depend on the bounding sequence $(M_n)_{n\in\mathbb{N}}$. There is a linear operator $\mathbb{I}(f)$ on $\mathcal{D}(\mathbb{I}(f))$ defined by

$$\text{(B.8)} \qquad \mathbb{I}(f)v = \lim_{n\to\infty} \mathbb{I}(f\chi_{M_n})v \quad \textit{for } v \in \mathcal{D}(\mathbb{I}(f)) \, .$$

(iii) $\bigcup_{n\in\mathbb{N}} \mathbf{E}(M_n)\mathcal{H}$ is contained in $\mathcal{D}(\mathbb{I}(f))$ and is a core for $\mathbb{I}(f)$. Moreover,

$$\text{(B.9)} \qquad \mathbf{E}(M_n)\mathbb{I}(f) \subset \mathbb{I}(f)\mathbf{E}(M_n) = \mathbb{I}(f\chi_{M_n}) \quad \textit{for } n \in \mathbb{N} \, .$$

Proof. For part (i), suppose that $v \in \mathcal{D}(\mathbb{I}(f))$. Since f is bounded on M_n one has $f\chi_{M_n}$ is bounded on Ω, and $\mathbb{I}(f\chi_{M_n})$ is everywhere defined and bounded (see [88, Section 4.3.1]). From [88, Proposition 4.12]

$$\begin{aligned}
\|\mathbb{I}(f\chi_{M_k})v - \mathbb{I}(f\chi_{M_n})\|_{\mathcal{H}}^2 &= \|\mathbb{I}(f\chi_{M_k} - f\chi_{M_n})\|_{\mathcal{H}}^2 \\
&= \int_\Omega |f\chi_{M_k} - f\chi_{M_n}|^2 \,\mathrm{d}\mu_v(t) \\
&= \|f\chi_{M_k} - f\chi_{M_n}\|_{L^2(\Omega,\mu_v(t))}
\end{aligned}$$

for $k, n \in \mathbb{N}$. Since $f \in L^2(\Omega, \mu_v(t))$, by (B.7), and the fact that $f\chi_{M_n} \to f$ in $L^2(\Omega, \mu_v(t))$ from Lebesgue dominated convergence (see [79, Theorem 1.34]), $(f\chi_{M_n})_{n\in\mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega, \mu_v(t))$. Thus $(\mathbb{I}(f\chi_{M_n})v)_{n\in\mathbb{N}}$ converges in $\mathcal{H}$, and is therefore the $\mathcal{H}$-norms are uniformly bounded in n. This completes the forward implication of part (i).

For the backward implication of part (i), by assumption

$$c := \sup_{n\in\mathbb{N}}\{\|\mathbb{I}(f\chi_{M_n})\|_{\mathcal{H}}\} < \infty \, .$$

Since $(|f\chi_{M_n}|^2)_{n\in\mathbb{N}}$ converges monotonically to $|f|^2$ $\mu_v(t)$-a.e. on Ω, by Lebesgue's monotone convergence theorem (see [79, Theorem 1.26])

$$\int_\Omega |f|^2 \,\mathrm{d}\mu_v(t) = \lim_{n\to\infty} \int_\Omega |f\chi_{M_n}|^2 \,\mathrm{d}\mu_v(t)$$

$$= \lim_{n\to\infty} \|\mathbb{I}\,(f\chi_{M_n})\|_{\mathcal{H}}^2 \le c^2 < \infty\,,$$

where the second equality is due to [88, Proposition 4.12]. As such, $f \in L^2(\Omega, \mu_v\,(t))$ and so $x \in \mathcal{D}\,(\mathbb{I}\,(f))$.

For part (ii), let $(M'_n)_{n\in\mathbb{N}}$ be another bounding sequence for f. From [88, Proposition 4.12] one has

$$\begin{aligned}\|\mathbb{I}\,(f\chi_{M_n})\,v\ -&\mathbb{I}\,(fM'_k)\,v\|_{\mathcal{H}} = \left\|f\chi_{M_n} - f\chi_{M'_k}\right\|_{L^2(\Omega,\mu_v(t))}\\ &\le \|f\chi_{M_n} - f\|_{L^2(\Omega,\mu_v(t))} + \left\|f - f\chi_{M'_k}\right\|_{L^2(\Omega,\mu_v(t))} \to 0\,,\end{aligned}$$

as $k, n \to \infty$ since $f\chi_{M_n} \to f$ and $f\chi_{M'_k} \to f$ in $L^2(\Omega, \mu_v\,(t))$ as noted in part (i). Therefore, $\lim_{n\to\infty} \mathbb{I}\,(f\chi_{M_n})\,v = \lim_{k\to\infty} \mathbb{I}\left(f\chi_{M'_k}\right) v$. This proves part (ii).

For part (iii), let $v \in \mathcal{H}$, and since $\mathbf{E}\,(M_k) = \mathbb{I}\,(\chi_{M_k})$, from [88, Proposition 4.12], one has

$$(*)\quad \mathbb{I}\,(f\chi_{M_k})\,v = \mathbb{I}\,(f\chi_{M_n}\chi_{M_k})\,v = \mathbb{I}\,(f\chi_{M_n})\,\mathbf{E}\,(M_k)\,v = \mathbf{E}\,(M_k)\,\mathbb{I}\,(f\chi_{M_n})\,v\,,$$

for $n \ge k$. So, $\sup_{n\in\mathbb{N}} \|\mathbb{I}\,(f\chi_{M_n})\,v\|_{\mathcal{H}} < \infty$, so that $\mathbf{E}\,(M_k)\,v \in \mathcal{D}\,(\mathbb{I}\,(f))$ by part (i). That is to say, $\bigcup_{k\in\mathbb{N}} \mathbf{E}\,(M_k)\,\mathcal{H} \subset \mathcal{D}\,(\mathbf{E}\,(f))$.

Now, taking $n \to \infty$ and using (B.8), one has that $\mathbb{I}\,(f)\,\mathbf{E}\,(M_k)\,v = \mathbb{I}\,(f\chi_{M_k})\,v$ for $v \in \mathcal{H}$. Now suppose $v \in \mathcal{D}\,(\mathbb{I}\,(f))$. Letting $n \to \infty$ again in (*), one obtains $\mathbf{E}\,(M_k)\,\mathbb{I}\,(f)\,v = \mathbb{I}\,(f)\,\mathbf{E}\,(M_k)\,v$. This proves (B.9).

Since $\mathbf{E}\,(M_n)\,v \to v$ and $\mathbb{I}\,(f)\,\mathbf{E}\,(M_n)\,v = \mathbf{E}\,(M_n)\,\mathbb{I}\,(f)\,v \to \mathbb{I}\,(f)\,v$ for $v \in \mathcal{D}\,(\mathbb{I}\,(f))$, then the linear subspace $\bigcup_{n\in\mathbb{N}} \mathbf{E}\,(M_n)\,\mathcal{H}$ of $\mathcal{H}$ is a core for $\mathbb{I}\,(f)$. This concludes the proof of part (iii). □

Theorem B.2.13 (Proposition 4.18 [88])**.** *The operator* $\mathbb{I}\,(f)$ *is bounded if and only if* $f \in L^\infty(\Omega, \mathbf{E})$*. Should this be the case, then* $\|\mathbb{I}\,(f)\|_{\text{op}} = \|f\|_{L^\infty(\Omega,\mathbf{E})}$.

Proof. From (B.6) it is immediate that if $f \in L^\infty(\Omega, \mathbf{E})$, then $\mathbb{I}\,(f)$ is bounded and $\|\mathbb{I}\,(f)\|_{\text{op}} \le \|f\|_{L^\infty(\Omega,\mathbf{E})}$.

Now suppose that $\mathbb{I}\,(f) \in \mathscr{B}(\mathcal{H})$. Set $M_n = \{t \in \mathbb{R};\ |f(t)| \ge \|\mathbb{I}\,(f)\|_{\text{op}} + 2^{-n}\}$ for $n \in \mathbb{N}$. By Theorem B.2.16 and (B.6) for $x \in \mathcal{H}$, one has

$$\|\mathbb{I}\,(f)\|_{\text{op}}^2\,\|\mathbf{E}\,(M_n)\,x\|_{\mathcal{H}}^2 \ge \|\mathbb{I}\,(f)\,\mathbf{E}\,(M_n)\,x\|_{\mathcal{H}}^2 = \|\mathbb{I}\,(f\chi_{M_n})\|_{\mathcal{H}}^2$$

$$\begin{aligned}&= \int_\Omega |f\chi_{M_n}|^2 \, \mathrm{d}\mu_x(t) = \int_{M_n} |f|^2 \, \mathrm{d}\mu_x(t) \\ &\geq (\|\mathbb{I}(f)\|_{\mathrm{op}} + 2^{-n})^2 \, \|\mathbf{E}(M_n)\, x\|_{\mathcal{H}}^2 \, .\end{aligned}$$

It is clear then that $\mathbf{E}(M_n)\, x = 0$, so $\mathbf{E}(M_n) = \mathbb{O}$ and therefore $\mathbf{E}\left(\bigcup_{n\in\mathbb{N}} M_n\right) = \mathbb{O}$. Since $M = \{t \in \mathbb{R};\, |f(t)| > \|\mathbb{I}(f)\|_{\mathrm{op}}\}$, this means that $|f(t)| \leq \|\mathbb{I}(f)\|_{\mathrm{op}}$ $\mathbf{E}$-a.e., so $\|f\|_{L^\infty(\Omega,\mathbf{E})} \leq \|\mathbb{I}(f)\|_{\mathrm{op}}$. □

It may well be the case that some spectral integrals give rise to unbounded operators. It is therefore critical that the domain of definition of these spectral integrals be specified.

Definition B.2.14. Suppose that f is an $\mathbf{E}$-a.e. finite measurable function on the space $(\Omega, \mathfrak{M})$. Then the domain of the spectral integral is

$$\mathcal{D}(\mathbb{I}(f)) := \left\{ x \in \mathcal{H};\, \int_\Omega |f(t)|^2 \, \mathrm{d}\mu_x(t) < \infty \right\} . \tag{B.10}$$

Proposition B.2.15 (Proposition 4.15 [88]). *Let $f, g \in \mathcal{S}$ and $x \in \mathcal{D}(\mathbb{I}(f))$ and $y \in \mathcal{D}(\mathbb{I}(g))$. Then*

$$\langle \mathbb{I}(g)\, y,\, \mathbb{I}(f)\, x\rangle = \int_\Omega f(t)\overline{g(t)} \, \mathrm{d}\mu_{x,y}(t) \, . \tag{B.11}$$

Proof. To begin, if f is a bounded $\mathfrak{M}$-measurable function on Ω, then one has that $\langle y,\, \mathbb{I}(f)\, x\rangle = \int_\Omega f(t) d\mu_{x,y}(t)$. This is easily seen by taking a simple function for f and noting that the bounded measurable functions are arbitrarily well-approximated by simple functions on Ω (see [88, Prop. 4.12] for details). Another useful couple of identities for this class of functions are $\mathbb{I}\left(\overline{f}\right) = \mathbb{I}(f)^*$ and $\mathbb{I}(fg) = \mathbb{I}(f)\, \mathbb{I}(g)$ (where g is also bounded measurable).

Now, consider the bounded function $f\overline{g}\chi_{M_n}$, where $(M_n)_{n\in\mathbb{N}}$ is a bounding sequence for f and g. From the comments above,

$$\int_\Omega f\overline{g}\chi_{M_n} \, \mathrm{d}\mu_{x,y}(t) = \langle y,\, \mathbb{I}(f\overline{g}\chi_{M_n})\, x\rangle = \langle \mathbb{I}(g\chi_{M_n})\, y,\, \mathbb{I}(f\chi_{M_n})\, x\rangle \, .$$

As $x \in \mathcal{D}(\mathbb{I}(f))$ and $y \in \mathcal{D}(\mathbb{I}(g))$, by Definition B.2.14 it means that $f \in L^2(\Omega, \mu_x(t))$ and $g \in L^2(\Omega, \mu_y(t))$. By Lemma B.2.8 the integral

$\int_\Omega f\overline{g}\,\mathrm{d}\mu_{x,y}(t)$ exists, and so

$$\begin{aligned}\left|\int_\Omega f\overline{g}\chi_{M_n}\,\mathrm{d}\mu_{x,y}(t)-\int_\Omega f\overline{g}\,\mathrm{d}\mu_{x,y}(t)\right|\\ &=\left|\int_\Omega (f\chi_{M_n}-f)\overline{g}\,\mathrm{d}\mu_{x,y}(t)\right|\\ &\leq \|f\chi_{M_n}-f\|_{L^2(\Omega,\mu_x(t))}\|g\|_{L^2(\Omega,\mu_y(t))}\xrightarrow{n\to\infty}0\,.\end{aligned}$$

Clearly, $f\chi_{M_n}\to f$ in the $L^2(\Omega,\mu_x(t))$ topology from application of the Lebesgue dominated convergence theorem [79, Theorem 1.34]. Furthermore, by Theorem B.2.12 (ii), one has that $\mathbb{I}(f)x=\lim_{n\to\infty}\mathbb{I}(f\chi_{M_n})x$ for any $x\in\mathcal{D}(\mathbb{I}(f))$. The claim has now been proven. □

Some of the most important properties of the spectral measure are listed below. For the proof, the reader is referred to [88].

Theorem B.2.16 (Theorem 4.16 [88])**.** *Let $f,g\in\mathcal{S}(\Omega,\mathfrak{M},\mathbf{E})$ and $\alpha,\beta\in\mathbb{C}$. Then one has*

(i) $\mathbb{I}\left(\overline{f}\right)=\mathbb{I}(f)^*$*, where $\overline{f}$ denotes the complex conjugate of the function f,*

(ii) $\mathbb{I}(\alpha f+\beta g)=\overline{\alpha\mathbb{I}(f)+\beta\mathbb{I}(g)}$*,*

(iii) $\mathbb{I}(fg)=\overline{\mathbb{I}(f)\mathbb{I}(g)}$*,*

(iv) $\mathbb{I}(f)$ *is a closed normal operator on $\mathcal{H}$, and* $\mathbb{I}(f)^*\mathbb{I}(f)=\mathbb{I}\left(\overline{f}f\right)=\mathbb{I}(f)\mathbb{I}(f)^*$

(v) $\mathcal{D}(\mathbb{I}(f)\mathbb{I}(g))=\mathcal{D}(\mathbb{I}(fg))\cap\mathcal{D}(\mathbb{I}(g))$.

Theorem B.2.17 (Proposition 4.17 [88])**.** *Let $f,g\in\mathcal{S}(\Omega,\mathfrak{M},\mathbf{E})$.*

(i) If $f(t)=g(t)$ $\mathbf{E}$-a.e. on Ω, then $\mathbb{I}(f)=\mathbb{I}(g)$.

(ii) If $f(t)$ is real $\mathbf{E}$-a.e. on Ω, then $\mathbb{I}(f)$ is self-adjoint.

(iii) If $f(t)\geq 0$ $\mathbf{E}$-a.e. on Ω, then $\mathbb{I}(f)$ is positive and self-adjoint.

The next two results for the spectral integral presented here are important in the construction of the functional calculus for self-adjoint operators.

Proposition B.2.18 (Proposition 4.19 [88]). *The spectral integral operator $\mathbb{I}\,(f)$ is invertible if and only if $f(t) \neq 0$ $\mathbf{E}$-a.e. on Ω. In this case, one has $\mathbb{I}\,(f)^{-1} = \mathbb{I}\,(f^{-1})$.*

Remark B.2.19. This is needed in the use of the functional calculus for the representation of the inverse of an injective, self-adjoint operator A, namely A^{-1}, in terms of a spectral integral.

The following proposition is particularly important for polynomial Krylov subspaces.

Proposition B.2.20 (Proposition 4.22 [88]). *For any polynomial p on $\mathbb{R}$ with coefficients in $\mathbb{C}$, one has $\mathbb{I}\,(p(f)) = p\,(\mathbb{I}\,(f))$, given $f \in \mathcal{S}(\Omega, \mathfrak{M}, \mathbf{E})$.*

Sketch of the proof. Induction is used on a general polynomial p of given degree n. Suppose that the assertion holds true for each polynomial of degree strictly less than n. Then, all one must do is prove the assertion remains true for the particular polynomial $p(t) = t^n$.

Take the identity $|f|^2 \leq 1 + |f^n|^2$, which is easily verified by considering where $|f(t)| \leq 1$ and $|f(t)| > 1$ separately. Then, by (B.10) $\mathcal{D}\,(\mathbb{I}\,(f^n)) \subset \mathcal{D}\,(\mathbb{I}\,(f))$, and using the induction hypothesis $\mathbb{I}\,(f^{n-1}) = \mathbb{I}\,(f)^{n-1}$ along with Theorem B.2.16 (v),

$$\begin{aligned}\mathcal{D}\,(\mathbb{I}\,(f)^n) &= \mathcal{D}\left(\mathbb{I}\,(f)^{n-1}\,\mathbb{I}\,(f)\right) = \mathcal{D}\left(\mathbb{I}\left(f^{n-1}\right)\mathbb{I}\,(f)\right)\\ &= \mathcal{D}\,(\mathbb{I}\,(f)) \cap \mathcal{D}\,(\mathbb{I}\,(f^n)) = \mathcal{D}\,(\mathbb{I}\,(f^n))\ .\end{aligned}$$

Now by Theorem B.2.16 (iii) it follows that

$$\mathbb{I}\,(f)^n = \mathbb{I}\,(f)^{n-1}\,\mathbb{I}\,(f) = \mathbb{I}\left(f^{n-1}\right)\mathbb{I}\,(f) \subset \mathbb{I}\,(f^n)\ ,$$

and using Theorem B.2.16 (ii), for a general polynomial $p(t)$ of degree n, one obtains

$$p\,(\mathbb{I}\,(f)) \subset \mathbb{I}\,(p(f))\ .$$

Furthermore, $f^n \in L^2(\Omega, \mu_x\,(t))$ if and only if $p(f) \in L^2(\Omega, \mu_x\,(t))$. So $\mathcal{D}\,(\mathbb{I}\,(f^n)) = \mathcal{D}\,(\mathbb{I}\,(p(f)))$ by (B.10). And also $\mathcal{D}\,(p\,(\mathbb{I}\,(f))) = \mathcal{D}\,(\mathbb{I}\,(f)^n)$ by the definition of $p\,(\mathbb{I}\,(f))$. The result then follows. □

The spectrum of the spectral integral operator is as follows.

Proposition B.2.21 (Proposition 4.20 [88]). *The spectrum of $\mathbb{I}(f)$ is the essential range of f, i.e.*

$$\sigma(\mathbb{I}(f)) = \{\zeta \in \mathbb{C};\, \mathbf{E}(\{t \in \Omega;\, |f(t) - \zeta| < \varepsilon\}) \neq \mathbb{O}, \quad \forall \varepsilon > 0\}\,. \tag{B.12}$$

Moreover, if $\zeta \in \rho(\mathbb{I}(f))$, then $\mathcal{R}(\mathbb{I}(f), \zeta) = \mathbb{I}((f - \zeta)^{-1})$.

Proof. Set the function $\widetilde{f} = f - \zeta$.

Now $0 \in \rho\left(\mathbb{I}\left(\widetilde{f}\right)\right)$ if and only if $\mathbb{I}\left(\widetilde{f}\right)$ has a bounded, everywhere defined inverse. Therefore, by Proposition B.2.18 and Theorem B.2.13 it follows that $\widetilde{f}^{-1} \in L^\infty(\Omega, \mathbf{E})$. Equivalently, there must exist some constant $c > 0$ such that $\mathbf{E}\left(\{t \in \mathbb{R};\, |\widetilde{f}(t)| \geq c\}\right) = \mathbb{O}$.

So, $0 \in \sigma\left(\mathbb{I}\left(\widetilde{f}\right)\right)$ if and only if $\mathbf{E}\left(\{t \in \mathbb{R};\, |\widetilde{f}(t)| < \varepsilon\}\right) \neq \mathbb{O}$ for all $\varepsilon > 0$ (precisely as the spectrum and resolvent sets are complementary). So by Proposition B.2.18 it follows that $\mathcal{R}(\mathbb{I}(f), \zeta) = \mathbb{I}(f - \zeta)^{-1} = \mathbb{I}((f - \zeta)^{-1})$. □

B.2.3 The spectral representation theorem

The general spectral theorem for unbounded self-adjoint operators is stated in what follows. This obviously simplifies in the case where one is dealing with a bounded operator. The definition of the functional calculus is then given. It is immediate from the definition of the functional calculus that it is simply a spectral integral, and hence all the theory of the previous section applies.

Theorem B.2.22 (Theorem 5.7 [88]). *Let A be a self-adjoint operator on a Hilbert space $\mathcal{H}$. Then there exists a unique spectral measure $\mathbf{E}$ on the Borel σ-algebra $\mathfrak{B}(\mathbb{R})$ such that*

$$A = \int_{\mathbb{R}} t \, \mathrm{d}\mathbf{E}(t)\,. \tag{B.13}$$

Remark B.2.23. It is immediate from Proposition B.2.21 and the definition of the support of the spectral measure, that the interval of support of the measure $\mathbf{E}$ in Theorem B.2.22 is in fact the *spectrum* of the operator A itself. Therefore, the interval of integration in B.13 may be replaced with the spectrum $\sigma(A)$.

Definition B.2.24. For a self-adjoint operator A on the Hilbert space $\mathcal{H}$, and $\mathbf{E}$ the unique spectral measure of Theorem B.2.22, the mapping that takes some $f \in \mathcal{S}(\mathbb{R}, \mathfrak{B}(\mathbb{R}), \mathbf{E})$ to the spectral integral $\mathbb{I}\,(f)$ is the *functional calculus* of the operator A. The functional calculus for the function f with respect to the operator A is denoted $f(A)$, and it has the same properties as the spectral integral.

B.3 The general functional calculus for the class $\mathscr{B}(\mathcal{H})$

In this Section, a few results are given for the functional calculus of general bounded operators on Hilbert space. Details may be found in [76, Chapter XI].

Before the statement of the theorem for the general functional calculus or spectral mapping, first one must define the concept of an admissible domain.

Definition B.3.1 (Chapter XI, Section 148 [76])**.** An admissible domain with respect to a linear operator A is any bounded open set $\mathcal{U}$ in $\mathbb{C}$, whose boundary $\partial\mathcal{U}$ consists of a finite number of rectifiable closed curves lying in the resolvent set $\rho(A)$, with the same orientation as $\mathcal{U}$ as a subset of the complex plane.

Theorem B.3.2 (Chapter XI, Section 151 [76])**.** *Let A be a bounded linear operator on Hilbert space $\mathcal{H}$. Consider the complex function $f(z)$ that is defined and differentiable, with respect to z, at all points of an open subset of $\mathcal{V} \subset \mathbb{C}$ that contains the spectrum of A. Let $\mathcal{U}$ be an admissible domain with respect to A containing the entire spectrum of A, and itself with its boundary* *contained in $\mathcal{V}$. Then the general functional calculus $f(A)$ is given by the following Cauchy integral.*

$$f(A) = \frac{1}{2\pi i} \int_{\partial \mathcal{U}} f(z) \mathcal{R}\,(A,\, z)\ \mathrm{d}z\,. \tag{B.14}$$

In this case, the following relation between the spectrum of the operator A and the functional calculus $f(A)$ holds.

$$\sigma(f(A)) = f(\sigma(A))\,. \tag{B.15}$$

Remark B.3.3. As the resolvent function $\mathcal{R}\,(A,\, z)$ is analytic in z on $\rho(A)$ [76, 51] then $\|\mathcal{R}\,(A,\, z)\|_{\mathrm{op}}$ is bounded on $\partial\mathcal{U} \subset \rho(A)$. As a result, if there is a sequence of holomorphic functions on $\mathcal{V}$, $(f_n)_{n\in\mathbb{N}}$, that tend to $f \in H(\mathcal{V})$ uniformly in the domain $\mathcal{V}$, then the transformations $(f_n(A))_{n\in\mathbb{N}}$ tend to the transform $f(A)$ in the operator norm topology [76, Chapter XI, Section 151].

Appendix C

Functional Analysis Miscellanea

This Appendix contains some miscellanea from functional analysis that do not quite fit in the other appendices, but nonetheless are important for several arguments within this thesis.

C.1 Types of convergence in Hilbert space

In infinite-dimensional Banach spaces, there are several different notions of convergence. Here, the focus is on three different forms of convergence in Hilbert space, namely strong convergence, weak convergence, and pointwise convergence.

Definition C.1.1. Let $\mathcal{H}$ be a Hilbert space, and let $(u_n)_{n\in\mathbb{N}}$ be a sequence of vectors in $\mathcal{H}$. Then

(i) The *strong convergence* (norm convergence) of $(u_n)_{n\in\mathbb{N}}$ to a vector $u \in \mathcal{H}$ is said to occur if $\|u_n - u\|_{\mathcal{H}} \xrightarrow{n\to\infty} 0$. It is customarily written $u_n \to u$ (or $u_n \xrightarrow{\|\cdot\|_{\mathcal{H}}} u$) as $n \to \infty$.

(ii) The *weak convergence* of $(u_n)_{n\in\mathbb{N}}$ to a vector $u \in \mathcal{H}$ is equivalent to $\langle v, u_n\rangle \to \langle v, u\rangle$ as $n \to \infty$ for all $v \in \mathcal{H}$ [10, Proposition 3.5]. Weak convergence is written $u_n \rightharpoonup u$ as $n \to \infty$.

(iii) For a separable Hilbert space $\mathcal{H}$ with an orthonormal basis $(e_k)_{k\in\mathbb{N}}$, the *component-wide convergence* of $(u_n)_{n\in\mathbb{N}}$ to a vector $u \in \mathcal{H}$ is said to occur if $\langle e_k,\, u_n\rangle \xrightarrow{n\to\infty} \langle e_k,\, u\rangle$ for all $k \in \mathbb{N}$. This means that the k-th component of the u_n's converge to the k-th component of u with respect to the basis $(e_k)_{k\in\mathbb{N}}$. Component-wise convergence is customarily written as $u_n \rightsquigarrow u$ as $n \to \infty$.

Remark C.1.2. The implications

$$\text{strong} \quad \Rightarrow \quad \text{weak} \quad \Rightarrow \quad \text{component-wise} \tag{C.1}$$

hold on $\mathcal{H}$, but their converses do not when $\dim \mathcal{H} = \infty$ (whereas the converses are true when $\dim \mathcal{H} < \infty$).

Lemma C.1.3. *Let $\mathcal{H}$ be a separable Hilbert space, and $(u_n)_{n\in\mathbb{N}}$ a sequence of vectors in $\mathcal{H}$. Then for $u \in \mathcal{H}$ one has the following.*

$$u_n \xrightarrow{n\to\infty} u \quad \Leftrightarrow \quad \begin{cases} u_n \rightharpoonup u \\ \|u_n\|_{\mathcal{H}} \to \|u\|_{\mathcal{H}} \end{cases}, \tag{C.2}$$

and

$$u_n \rightharpoonup u \text{ as } n \to \infty \quad \Leftrightarrow \quad \begin{cases} \langle e_k,\, u_n\rangle \xrightarrow{n\to\infty} \langle e_k,\, u\rangle \;\; \forall k \in \mathbb{N} \\ \sup_{n\in\mathbb{N}} \|u_n\|_{\mathcal{H}} < \infty \end{cases}, \tag{C.3}$$

where $(e_k)_{k\in\mathbb{N}}$ is an orthonormal basis of $\mathcal{H}$.

Proof. Obviously the forward implication of (C.2) holds [10, Proposition 3.5]. For the backward implication note that $\|u_n - u\|^2_{\mathcal{H}} = \langle u_n - u,\, u_n - u\rangle$ so that

$$\|u_n - u\|^2_{\mathcal{H}} = \|u_n\|^2_{\mathcal{H}} - \langle u,\, u_n\rangle - \langle u_n,\, u\rangle + \|u\|^2_{\mathcal{H}} \,.$$

From the weak convergence $u_n \rightharpoonup u$, one has $\langle u,\, u_n\rangle \xrightarrow{n\to\infty} \|u\|^2_{\mathcal{H}}$ and $\langle u_n,\, u\rangle \xrightarrow{n\to\infty} \|u\|^2_{\mathcal{H}}$. The result then follows.

To show the forward implication in (C.3) is a straightforward application of the definition of weak convergence and the fact that under weak convergence one has that $(\|u_n\|_{\mathcal{H}})_{n\in\mathbb{N}}$ is uniformly bounded [10, Proposition 3.5].

For the backward implication of (C.3), consider some $f = \sum_{k\in\mathbb{N}} f_k e_k \in \mathcal{H}$. Then

$$\langle f,\, u_n - u\rangle = \sum_{k\in\mathbb{N}} \overline{f_k}\,\langle e_k,\, u_n - u\rangle$$

and for some fixed $M \in \mathbb{N}$

$$\begin{aligned}
|\langle f,\, u_n - u\rangle| &\leq \sum_{k=1}^{M} |f_k|\,|\langle e_k,\, u_n - u\rangle| + \sum_{k=1+M}^{\infty} |f_k|\,|\langle e_k,\, u_n - u\rangle| \\
&\leq \max_{k\in\{1,\dots,M\}} |f_k|\,|\langle e_k,\, u_n - u\rangle| + \sum_{k=1+M}^{\infty} |f_k|\,|\langle e_k,\, u_n - u\rangle| \\
&\leq \max_{k\in\{1,\dots,M\}} |f_k|\,|\langle e_k,\, u_n - u\rangle| \\
&\quad + \left(\sum_{k=1+M}^{\infty} |f_k|^2\right)^{\frac{1}{2}} \left(\sum_{k=1+M}^{\infty} |\langle e_k,\, u_n - u\rangle|^2\right)^{\frac{1}{2}} \\
&\leq \max_{k\in\{1,\dots,M\}} |f_k|\,|\langle e_k,\, u_n - u\rangle| + \|u_n - u\|_{\mathcal{H}}^2 \left(\sum_{k=1+M}^{\infty} |f_k|^2\right)^{\frac{1}{2}} .
\end{aligned}$$

As $\|u_n - u\|_{\mathcal{H}}$ is uniformly bounded in n and $u_n \rightsquigarrow u$, the above inequality may be arbitrarily small by choosing both $M \in \mathbb{N}$ and $n \in \mathbb{N}$ large enough. The result then follows. □

For convenience, below are the definitions of some different types of operator convergence.

Definition C.1.4. Consider a family of linear operators in Hilbert space $A_n : \mathcal{H} \to \mathcal{H}$ such that $A_n \in \mathscr{B}(\mathcal{H})$ for all $n \in \mathbb{N}$.

(i) If there exists some $A \in \mathscr{B}(\mathcal{H})$ such that $\|A_n - A\|_{\mathrm{op}} \xrightarrow{n\infty} 0$, then the A_n's are said to *converge in the operator norm.*

(ii) If there exists some $A \in \mathscr{B}(\mathcal{H})$ such that for every $\psi \in \mathcal{H}$ one has $\|(A_n - A)\psi\|_{\mathcal{H}} \xrightarrow{n\to\infty} 0$, i.e. $A_n\psi \to A\psi$, then the A_n's are said to *converge in the strong operator topology.*

(iii) If there exists some $A \in \mathscr{B}(\mathcal{H})$ such that for every $\psi \in \mathcal{H}$ one has

$A_n\psi \rightharpoonup A\psi$ as $n \to \infty$, then the A_n's are said to *converge in the weak operator topology.*

C.2 Some approximation theorems

One of the deepest results in approximation theory is the Stone-Weierstrass theorem. This is stated below for compact Hausdorff spaces, along with its generalisation to *locally* compact Hausdorff spaces. The full details and proofs may be found in [91]. To begin with, the precise definitions concerning the separation of points and vanishing are required.

Definition C.2.1. Let X be a space and let η be a subset of the functions from X to $\mathbb{C}$. Then η is said to vanish at a point $x_0 \in X$ if $f(x_0) = 0$ for all $f \in \eta$.

Definition C.2.2. Let X be a space and let η be a subset of the functions from X to $\mathbb{C}$. Then η is said to *separate points* on X if for every *distinct* x and y in X, there is some function $f \in \eta$ such that $f(x) \neq f(y)$.

Theorem C.2.3. *Let X be a compact Hausdorff space, and let $C(X, \mathbb{C})$ denote the space of continuous functions on X equipped with the supremum norm. If C is an involutive subalgebra of $C(X, \mathbb{C})$ that separates points in X, then either*

(i) the closure of C in the supremum norm is all of $C(X, \mathbb{C})$, or

(ii) the closure of C in the supremum norm is the family of all functions $\eta \subset C(X, \mathbb{C})$ that vanish at a uniquely determined point $x_0 \in X$.

In particular, if C is a unital, involutive subalgebra of $C(X, \mathbb{C})$ that separates points, then the closure of C in the supremum norm is all of $C(X, \mathbb{C})$, i.e. $\overline{C}^{\|\cdot\|_\infty} = C(X, \mathbb{C})$.

Proof. See [91, Section 5, Corollaries 1 & 2]. □

Theorem C.2.4. *Let X be a locally compact Hausdorff space, and let $C_0(X,\mathbb{C})$ denote the space of continuous functions on X that vanish at infinity equipped with the supremum norm. If C is a involutive subalgebra of $C_0(X,\mathbb{C})$ that separates points and vanishes nowhere, then the closure of C in the supremum norm is all of $C_0(X,\mathbb{C})$, i.e. $\overline{C}^{\|\cdot\|_\infty} = C_0(X,\mathbb{C})$.*

Proof. see [91, Section 6, Corollary 1]. □

The following classical result specifically concerns *polynomial* approximations to holomorphic functions.

Theorem C.2.5 (Theorem 13.7 [79])**.** *Let $K \subset \mathbb{C}$ be a compact set, and let $\mathbb{C}^* \setminus K$ be connected (where $\mathbb{C}^*$ denotes the single point compactification of $\mathbb{C}$). Let f be holomorphic in $\mathcal{U}$, where $\mathcal{U}$ is an open set containing K. Then there is a sequence of polynomials $(p_n)_{n\in\mathbb{N}}$ that approach $f(z)$ uniformly on K.*

Bibliography

[1] M. Abramowitz and I.A. Stegun. *Handbook of Mathematical Functions: With Formulas, Graphs, and Mathematical Tables.* Applied mathematics series. U.S. Department of Commerce, National Bureau of Standards, 1972.

[2] Nenad Antonić, Marko Erceg, and Alessandro Michelangeli. "Friedrichs systems in a Hilbert space framework: Solvability and multiplicity". In: *Journal of Differential Equations* 263 (2017), pp. 8264–8294.

[3] Nenad Antonić et al. "Complex Friedrichs systems and applications". In: *Journal of Mathematical Physics* 58.10 (2017), p. 101508.

[4] M. Arioli, V. Pták, and Z. Strakoš. "Krylov sequences of maximal length and convergence of GMRES". In: *BIT Numerical Mathematics* 38.4 (1998), pp. 636–643.

[5] W. E. Arnoldi. "The principle of minimized iterations in the solution of the matrix eigenvalue problem". In: *Quarterly of Applied Mathematics* 9.1 (1951), pp. 17–29.

[6] B Beauzamy. *Introduction to Operator Theory and Invariant Subspaces.* North-Holland Mathematical Library. Amsterdam: Elsevier Science Publishers B.V., 1988.

[7] Bernhard Beckermann and Lothar Reichel. "Error estimates and evaluation of matrix functions using the Faber transorm". In: *SIAM Journal on Numerical Analysis* 47.5 (2009), pp. 3849–3883.

[8] Carl M Bender and Steven A Orszag. *Advanced Mathematical Methods for Scientists and Engineers I: Asymptotic Methods and Perturbation Theory*. New York: Springer-Verlag, 1999.

[9] Helmut Brakhage. "On ill-posed problems and the method of conjugate gradients". In: *Inverse and Ill-Posed Problems*. Ed. by Heinz W. Engl and C.W. Groetsch. Academic Press, 1987, pp. 165–175.

[10] Haim Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. first. New York: Springer, 2011.

[11] Peter N. Brown and Homer F. Walker. "GMRES on (nearly) singular systems". In: *SIAM Journal on Matrix Analysis and Applications* 18.1 (1997), pp. 37–51.

[12] S. L. Campbell et al. "GMRES and the minimal polynomial". In: *BIT Numerical Mathematics* 36.4 (1996), pp. 664–675.

[13] S.L. Campbell et al. "Convergence estimates for solution of integral equations with GMRES". In: *The Journal of Integral Equations and Applications* 8.1 (1996), pp. 19–34.

[14] Noe Caruso and Alessandro Michelangeli. *Convergence of the conjugate gradient method with unbounded operators*. preprint (2019), arXiv:1908.10110.

[15] Noe Caruso and Alessandro Michelangeli. *Krylov solvability of unbounded inverse linear problems*. SISSA preprint 25/2019/MATE (2019).

[16] Noe Caruso, Alessandro Michelangeli, and Paolo Novati. "On Krylov solutions to infinite-dimensional inverse linear problems". In: *Calcolo* 56.3 (2019), p. 32.

[17] Noe Caruso, Alessandro Michelangeli, and Paolo Novati. *Truncation and convergence issues for bounded linear inverse problems in Hilbert space*. preprint (2018), arXiv:1811.08195.

[18] Noè A. Caruso et al. "Spontaneous morphing of equibiaxially prestretched elastic bilayers: The role of sample geometry". In: *International Journal of Mechanical Sciences* 149 (2018), pp. 481 –486.

[19] T.S. Chihara. *An Introduction to Orthogonal Polynomials.* New York: Dover Publications, 1978.

[20] P. Ciarlet. *The Finite Element Method for Elliptic Problems.* Society for Industrial and Applied Mathematics, 2002.

[21] James W. Daniel. “The conjugate gradient method for linear and nonlinear operator equations”. In: *SIAM Journal on Numerical Analysis* 4 (1967), pp. 10–26.

[22] J Dongarra and F Sullivan. “The Top 10 Algorithms (Guest editors’ intruduction)”. In: *Computing in Science and Engineering* 2 (2000), pp. 22–23.

[23] Vladimir Druskin, Leonid Knizhnerman, and Mikhail Zaslavsky. “Solution of large scale evolutionary problems using rational Krylov supspaces with optimized shifts”. In: *SIAM Journal on Scientific Computing* 31.5 (2009), pp. 3760–3780.

[24] Bertolt Eicke, Alfred K. Louis, and Robert Plato. “The instability of some gradient methods for ill-posed problems”. In: *Numerische Mathematik* 58.1 (1990), pp. 129–134.

[25] Heinz W. Engl, Martin Hanke, and Andreas Neubauer. *Regularization of inverse problems.* Vol. 375. Mathematics and its Applications. Kluwer Academic Publishers Group, Dordrecht, 1996, pp. viii+321.

[26] Thomas Ericsson and Axel Ruhe. “The spectral transformation Lanczos method for the numerical solution of large sparse generalized symmetric eigenvalue problems”. In: *Mathematics of Computation* 35.152 (1980), pp. 1251–1268.

[27] A Ern and JL Guermond. “Discontinuous Galerkin methods for Friedrichs’ ststems. I. General theory”. In: *SIAM Journal on Numerical Analysis* 44.2 (2006), pp. 753–778.

[28] Alexandre Ern and Jean-Luc Guermond. *Theory and Practice of Finite Elements.* Vol. 159. Applied Mathematical Sciences. New York: Springer-Verlag, 2004, pp. xiv+524.

[29] Alexandre Ern, Jean-Luc Guermond, and Gilbert Caplain. “An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems”. In: *Communications in Patrial Differential Equations* 32 (2007), pp. 317–341.

[30] Roland W. Freund and Marlis Hochbruck. “On the use of two QMR algorithms for solving singular systems and applications in Markov chain modeling”. In: *Numerical Linear Algebra with Applications* 1.4 (1994), pp. 403–420.

[31] E Gallopoulous and Y Saad. “Efficient solution of parabolic equations by Krylov approximation methods”. In: *SIAM Journal on Scientific and Statistical Computing* 13.5 (1992), pp. 1236–1264.

[32] M. G. Gasparo, A. Papini, and A. Pasquali. “Some properties of GMRES in Hilbert spaces”. In: *Numerical Functional Analysis and Optimization* 29.11-12 (2008), pp. 1276–1285.

[33] Silvia Gazzola, Paolo Novati, and Maria Rosaria Russo. “On Krylov projection methods and Tikhonov regularization”. In: *Electronic Transactions on Numerical Analysis* 44 (2015), pp. 83–123.

[34] L. Gehér. “Cyclic vectors of a cyclic operator span the space”. In: *Proceedings of the American Mathematical Society* 33 (1972), pp. 109–110.

[35] M. Gilles and A. Townsend. “Continuous Analogues of Krylov Subspace Methods for Differential Operators”. In: *SIAM Journal on Numerical Analysis* 57.2 (2019), pp. 899–924.

[36] G Godefroy and JH Shapiro. “Operators with Dense, Invariant, Cyclic Vector Manifolds”. In: *Journal of Functional Analysis* 98.2 (1991), pp. 229–269.

[37] A. Greenbaum, V. Pták, and Z. Strakoš. “Any Nonincreasing Convergence Curve is Possible for GMRES”. In: *SIAM Journal on Matrix Analysis and Applications* 17.3 (1996), pp. 465–469.

[38] Anne Greenbaum and Zdenek Strakos. "Matrices that Generate the same Krylov Residual Spaces". In: *Recent Advances in Iterative Methods.* Ed. by Gene Golub, Mitchell Luskin, and Anne Greenbaum. New York, NY: Springer New York, 1994, pp. 95–118.

[39] S Güttel. "Rational Krylov Methods for Operator Functions". PhD thesis. TU Bergakademie Freiberg, 2010.

[40] Paul Richard Halmos. *A Hilbert space problem book.* Second. Vol. 19. Graduate Texts in Mathematics. Encyclopedia of Mathematics and its Applications, 17. New York-Berlin: Springer-Verlag, 1982.

[41] Martin Hanke. *Conjugate gradient type methods for ill-posed problems.* Vol. 327. Pitman Research Notes in Mathematics Series. Longman Scientific & Technical, Harlow, 1995, pp. iv+134.

[42] Per Christian Hansen. *Rank-deficient and discrete ill-posed problems.* SIAM Monographs on Mathematical Modeling and Computation. Numerical aspects of linear inversion. Philadelphia, PA: Society for Industrial and Applied Mathematics (SIAM), 1998, pp. xvi+247.

[43] K Henrik, A Olsson, and Axel Ruhe. "Rational Krylov for eigenvalue computation and model order reduction". In: *BIT Numerical Mathematics* 46 (2006), pp. 99–111.

[44] Roland Herzog and Ekkehard Sachs. "Superlinear convergence of Krylov subspace methods for self-adjoint problems in Hilbert space". In: *SIAM Journal on Numerical Analysis* 53.3 (2015), pp. 1304–1324. ISSN: 0036-1429.

[45] Magnus R. Hestenes and Eduard Stiefel. "Methods of conjugate gradients for solving linear systems". In: *Journal of Research of the National Bureau of Standards* 49 (1952), pp. 409–436.

[46] M. Hochbruck and C. Lubich. "Error Analysis of Krylov Methods In a Nutshell". In: *SIAM Journal on Scientific Computing* 19.2 (1998), pp. 695–701.

[47] Jack P Holman. *Heat Transfer.* 10th. McGraw-Hill Series in Mechanical Engineering. New York: McGraw-Hill, 2010.

[48] Carl Jagels and Lothar Reichel. “The extended Krylov subspace methods and orthogonal Laurent polynomials”. In: *Linear Algebra and its Applications* 431.3–4 (2009), pp. 441–458.

[49] W. J. Kammerer and M. Z. Nashed. “On the convergence of the conjugate gradient method for singular linear operator equations”. In: *SIAM Journal on Numerical Analysis* 9 (1972), pp. 165–181.

[50] W. Karush. “Convergence of a method of solving linear problems”. In: *Proceedings of the American Mathematical Society* 3 (1952), pp. 839–851.

[51] Tosio Kato. *Perturbation Theory for Linear Operators.* Classics in Mathematics. Reprint of the 1980 edition. Springer-Verlag, Berlin, 1995, pp. xxii+619. ISBN: 3-540-58661-X.

[52] Tom H. Koornwinder. “Orthogonal polynomials”. In: *Computer algebra in quantum field theory.* Texts Monogr. Symbol. Comput. Springer, Vienna, 2013, pp. 145–170.

[53] M. A. Krasnosel’skii et al. *Approximate Solution of Operator Equations.* first. Groningen: Wolters-Noordhoff, 1972.

[54] Cornelius Lanczos. “An iteration methods for the solution of the eigenvalue problem of linear differential and integral operators”. In: *Journal of Research of the National Bureau of Standards* 45.4 (1950), pp. 255–282.

[55] Jörg Liesen and Zdeněk Strakoš. *Krylov subspace methods.* Numerical Mathematics and Scientific Computation. Principles and analysis. Oxford: Oxford University Press, 2013, pp. xvi+391.

[56] Alfred K. Louis. “Convergence of the conjugate gradient method for compact operators”. In: *Inverse and Ill-Posed Problems.* Ed. by Heinz W. Engl and C.W. Groetsch. Academic Press, 1987, pp. 177–183.

[57] I. Moret. “A Note on the Superlinear Convergence of GMRES”. In: *SIAM Journal on Numerical Analysis* 34.2 (1997), pp. 513–516.

[58] I Moret and P Novati. “RD-Rational approximations of the matrix exponential”. In: *BIT Numerical Mathematics* 44 (2004), pp. 595–615.

[59] James Munkres. *Topology.* second. Edinburgh: Pearson, 2014.

[60] Yasuki Nakayama. *Introduction to Fluid Mechanics.* Oxford: Butterworth-Heinemann, 2000.

[61] M. Z. Nashed. "Iterative methods for the solutions of a class of nonlinear operator equations". In: *Mathematics of Computation* 19.89 (1965), pp. 14–24.

[62] Claudia Negulescu. "Numerical analysis of a multiscale finite element scheme for the resolution of the stationary Schrödinger equation". In: *Numerische Mathematik* 108.4 (2008), pp. 625–652.

[63] A. S. Nemirovskiĭ. "Regularizing properties of the conjugate gradient method in ill-posed problems". In: *Akademiya Nauk SSSR. Zhurnal Vychislitel'noĭ Matematiki i Matematicheskoĭ Fiziki* 26.3 (1986), pp. 332–347, 477.

[64] A. S. Nemirovskiy and B. T. Polyak. "Iterative methods for solving linear ill-posed problems under precise information. I". In: *Izvestiya Akademii Nauk SSSR. Tekhnicheskaya Kibernetika* 2 (1984), pp. 13–25, 203.

[65] A. S. Nemirovskiy and B. T. Polyak. "Iterative methods for solving linear ill-posed problems under precise information. II". In: *Engineering Cybernetics* 22 (1984), pp. 50–57.

[66] Olavi Nevanlinna. *Convergence of Iterations for Linear Equations.* first. Basel: Springer, 1993.

[67] P. Novati. "A convergence result for some Krylov-Tikhonov methods in Hilbert spaces". In: 39.6 (2018), pp. 655–666.

[68] P. Novati. "Some Properties of the Arnoldi-Based Methods for Linear Ill-Posed Problems". In: *SIAM Journal on Numerical Analysis* 55.3 (2017), pp. 1437–1455.

[69] S. Olver. "GMRES for the Differentiation Operator". In: *SIAM Journal on Numerical Analysis* 47.5 (2009), pp. 3359–3373.

[70] Sheehan Olver. “Fast, numerically stable computation of oscillatory integrals with stationary points”. In: *BIT Numerical Mathematics* 50.1 (2010), pp. 149–171.

[71] Sheehan Olver. “Shifted GMRES for oscillatory integrals”. In: *Numerische Mathematik* 114.4 (2010), pp. 607–628.

[72] C. Paige and M. Saunders. “Solution of Sparse Indefinite Systems of Linear Equations”. In: *SIAM Journal on Numerical Analysis* 12.4 (1975), pp. 617–629.

[73] C. Paige and Z. Strakos. “Residual and Backward Error Bounds in Minimum Residual Krylov Subspace Methods”. In: *SIAM Journal on Scientific Computing* 23.6 (2002), pp. 1898–1923.

[74] Alfio Quarteroni. *Numerical models for differential problems.* Vol. 16. MS&A. Modeling, Simulation and Applications. Third edition. Springer, Cham, 2017, pp. xvii+681.

[75] Alfio M. Quarteroni and Alberto Valli. *Numerical Approximation of Partial Differential Equations.* 1st ed. 1994. 2nd printing. Springer Publishing Company, Incorporated, 2008. ISBN: 3540852670, 9783540852674.

[76] Frigyes Riesz and Béla Sz.-Nagy. *Functional analysis.* Translated by Leo F. Boron. Frederick Ungar Publishing Co., New York, 1955.

[77] J.R. Ringrose. *Compact Non-self-adjoint Operators.* Van Nostrand Reinhold mathematical studies. Van Nostrand Reinhold Company, 1971.

[78] H. L. Royden. *Real Analysis.* third. New Jersey: Prentice-Hall, 1998.

[79] Walter Rudin. *Real and Complex Analysis.* third. New York, NY, USA: McGraw-Hill, Inc., 1987. ISBN: 0070542341.

[80] Axel Ruhe. “Rational Krylov Algorithms for Nonsymmetric Eigenvalue Problems”. In: *Recent Advances in Iterative Methods.* Ed. by Gene Golub, Mitchell Luskin, and Anne Greenbaum. New York, NY: Springer New York, 1994, pp. 149–164.

[81] Axel Ruhe. “Rational Krylov algorithms for nonsymmetric eigenvalue problems. II: matrix pairs”. In: *Linear Algebra and its Applications* 198 (1994), pp. 283–295.

[82] Axel Ruhe. “Rational Krylov sequence methods for eigenvalue computation”. In: *Linear Algebra and its Applications* 58 (1984), pp. 391–405.

[83] Axel Ruhe. “The rational Krylov algorithm for nonsymmetric eigenvalue problems. III: complex shifts for real matrices”. In: *BIT Numerical Mathematics* 34 (1994), pp. 165–176.

[84] Axel Ruhe. “The two-sided arnoldi algorithm for nonsymmetric eigenvalue problems”. In: *Matrix Pencils*. Ed. by Bo Kågström and Axel Ruhe. Berlin, Heidelberg: Springer Berlin Heidelberg, 1983, pp. 104–120.

[85] Y. Saad. “Krylov subspace methods for solving large unsymmetric linear systems”. In: *Mathematics of Computation* 37.155 (1981), pp. 105–126.

[86] Y. Saad and M. Schultz. “GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems”. In: *SIAM Journal on Scientific and Statistical Computing* 7.3 (1986), pp. 856–869.

[87] Yousef Saad. *Iterative methods for sparse linear systems*. second. Philadelphia, PA: Society for Industrial and Applied Mathematics, 2003.

[88] Konrad Schmüdgen. *Unbounded self-adjoint operators on Hilbert space*. Vol. 265. Graduate Texts in Mathematics. Dordrecht: Springer, 2012, pp. xx+432. ISBN: 978-94-007-4752-4. DOI: 10.1007/978-94-007-4753-1.

[89] J. A. Shohat and J. D. Tamarkin. *The Problem of Moments*. American Mathematical Society Mathematical surveys, vol. I. American Mathematical Society, New York, 1943, pp. xiv+140.

[90] Valeria Simoncini and Daniel B. Szyld. “Recent computational developments in Krylov subspace methods for linear systems”. In: *Numerical Linear Algebra with Applications* 14.1 (2007), pp. 1–59.

[91] MH Stone. "A generalized Weierstrass approximation theorem". In: *Studies in Modern Analysis*. Ed. by RC Buck. Vol. 1. Studies in Mathematics. The Mathematical Association of America, 1962, pp. 30–87.

[92] Gábor Szegő. *Orthogonal polynomials*. Fourth. American Mathematical Society, Colloquium Publications, Vol. XXIII. American Mathematical Society, Providence, R.I., 1975, pp. xiii+432.

[93] Andrey N. Tikhonov and Vasiliy Y. Arsenin. *Solutions of ill-posed problems*. Translated from the Russian, Preface by translation editor Fritz John, Scripta Series in Mathematics. New York-Toronto: John Wiley & Sons, 1977, pp. xiii+258.

[94] Jasper Van Den Eshof and Marlis Hochbruck. "Preconditioning Lanczos approximations to the matrix exponential". In: *SIAM Journal on Scientific Computing* 27.4 (2006), pp. 1438–1457.

[95] Ferdinand Verhulst. *Methods and Applications of Singular Perturbations: Boundary Layers and Multiple Timescale Dynamics*. Vol. 50. Texts in Applied Mathematics. New York: Springer-Verlag, 2005.

[96] H.A. Van der Vorst and C. Vuik. "The superlinear convergence behaviour of GMRES". In: *Journal of Computational and Applied Mathematics* 48.3 (1993), pp. 327–341.

[97] Ragnar Winther. "Some superlinear convergence results for the conjugate gradient method". In: *SIAM Journal on Numerical Analysis* 17.1 (1980), pp. 14–17.