COMPACT LINEAR OPERATORS AND KRYLOV SUBSPACE METHODS

Jan Marthedal Rasmussen

LYNGBY 2001
IMM–EKS–2001–9

Preface

This report constitutes my Master of Science Thesis, written during the period from August 1st, 2000 to January 31st, 2001, at Informatics and Mathematical Modelling (IMM) and the Department of Mathematics (MAT), Technical University of Denmark.

My supervisors have been Professor Per Christian Hansen, IMM, and Associate Professor Michael Pedersen, MAT. I would like to thank both for their contagious enthusiasm and for always having the time for yet another question.

I would also like to thank Jesper Grooss and my big brother Thomas for good advice and for helping me proof-reading.

Finally, I thank my dear girlfriend Mette Kjærsig for her love and support.

Kgs. Lyngby, February 1, 2001

Jan Marthedal Rasmussen

Abstract

This thesis deals with linear ill-posed problems related to compact operators, and iterative Krylov subspace methods for solving discretized versions of these.

Linear compact operators in infinite dimensional Hilbert spaces will be investigated and several results on the singular values and eigenvalues for such will be presented. A large subset of linear compact operators consists of integral operators and many results will be based on the kernel of such operators.

Finite dimensional approximations to these operators will be considered by using Galerkin discretization. Several results will be shown stating how singular values and eigenvalues (and corresponding vectors) of infinite dimensional operators and their finite dimensional approximations are related.

Krylov subspace methods, with focus on GMRES, will be investigated in relation to discrete ill-posed problems, that is, linear finite dimensional systems of equations that originate from ill-posed problems. By using the spectral decomposition of the coefficient matrix, results on the convergence of GMRES are derived.

Keywords: *Compact linear operator, Krylov subspace method, eigenvalues, singular values.*

Resumé

Denne afhandling omhandler lineære “ill-posed” problemer relateret til kompakte operatorer, og iterative Krylov underrums metoder til at løse diskretiserede udgaver af disse.

Lineære kompakte operatorer i Hilbert rum af uendelig dimension vil blive undersøgt og flere resultater vedrørende singulære værdier og egenværdier af sådanne vil blive præsenteret. En stor delmængde af lineære kompakte operatorer udgøres af integraloperatorer og mange resultater vil blive baseret på kernen af sådanne operatorer.

Endelig-dimensionale tilnærmelser til disse operatorer vil blive betragtet ved brug af Galerkin diskretisering. Adskellige resultater vil blive vist omhandlende hvorledes singulære værdier og egenværdier (og tilhørende vektorer) af uendelig-dimensionale operatorer and deres endelig-dimensionale tilnærmelser er relaterede.

Krylov underrums metoder, med fokus på GMRES, vil blive undersøgt i relation til diskrete ill-posed problemer, dvs. linære endeligt-dimensionale ligningssystemer der kommer fra ill-posed problemer. Ved at benytte diagonalisering af koefficientmatricen vil resultater omkring GMRES’ konvergens blive udledt.

Nøgleord: *Kompakte lineære operatorer, Krylov underrums metoder, egenværdier, singulære værdier.*

Notation

If not explicitly stated otherwise, the following symbols and notation will be used throughout the thesis.

Boldface and upper case letters will be used for matrices, such as $\mathbf{A}$ or $\mathbf{X}$. Matrices with an integer subscript represents a column of the matrix, e.g. $\mathbf{A}_4$ means the 4th column of $\mathbf{A}$. A specific element of a matrix is represented by a double integer subscript, e.g. $\mathbf{A}_{2,5}$ means the element at row 2 and column 5 of $\mathbf{A}$.

Vectors (finite dimensional coordinate vectors) will also be shown in boldface but with lower case letters. An element of a vector will be referenced by an integer subscript indicating the element number such as $\mathbf{v}_2$. A subvector can be extracted by using $\mathbf{v}_{a:b} = [\mathbf{v}_a, \mathbf{v}_{a+1}, \ldots, \mathbf{v}_b]^T$. Note that vectors always are column vectors such that $\mathbf{A}\mathbf{v}$ and $\mathbf{v}^T\mathbf{A}$ make sense while $\mathbf{A}\mathbf{v}^T$ and $\mathbf{v}\mathbf{A}$ do not.

Symbols with an integer superscript in parentheses will be used as enumeration, e.g. $f^{(2)}$, $\mathbf{v}^{(3)}$ or $\mathbf{A}^{(4)}$. They will be used for symbols with commen properties, typically in connection with iterative methods.

Expressions such as $x^\star = \operatorname{argmin}_x F(x)$ means that the quantity $x^\star$ fulfills $F(x^\star) = \min_x F(x)$.

Notation	Meaning	Ref.	Page
$B(\mathcal{V})$	Equivalent to $B(\mathcal{V},\mathcal{V})$	Sec. 2.3	14
$B(\mathcal{V},\mathcal{W})$	The set of linear and bounded operators from $\mathcal{V}$ into $\mathcal{W}$	Sec. 2.3	14
$C(I)$	The set of continuous functions on I	-	46
$\mathrm{diag}(\mathbf{x})$	Diagonal matrix with the vector $\mathbf{x}$ along the diagonal	Eq. (6.6)	78
$\dim(\mathcal{V})$	The dimension of the vector space $\mathcal{V}$	Sec. 7.3	96
$\mathcal{D}(K)$	The domain of K	Eq. (4.2)	38
$\mathbf{A}^H$	The Hermitian (or conjugate transposed) of $\mathbf{A}$	–	77
$\mathcal{H}$	Hilbert space	Sec. 2.4	15
I	Identity operator	Eq. (3.1)	26
I	Real interval	Sec. 4.1	37
$\mathbf{I}$	Identity matrix	Eq. (7.2)	93
$k(s,t)$	Integral kernel, $k \in L^2(I \times J)$	Eq. (4.1)	37
K	Compact linear operator	Def. 2.14	19
$\widetilde{K}_{MN}, \widetilde{K}_N$	Finite dimensional approximations to K	Eq. (6.3)	75
$\mathcal{K}_n(T,x)$	Krylov subspace	Def. 7.1	89
$\mathcal{N}(T)$	The null-space of T, $\mathcal{N}(T) = \{x \mid Tx = 0\}$	Eq. (2.12)	19
$\mathcal{O}(\cdot)$	If $f(n) = \mathcal{O}(g(n))$ then $f(n) \leq Cg(n)$ for all n for some constant $C > 0$	Eq. (3.14)	34
$\mathcal{P}_k$	The set of polynomials p of maximum degree k and for which $p(0) = 1$	–	98
$\mathcal{R}(K)$	The range of K, $\mathcal{R}(K) = K(\mathcal{D}(K))$	–	38
T	Linear operator	Sec. 2.3	14
$\mathbf{A}^T$	The transposed of $\mathbf{A}$	–	–
u_n	Eigenfunction of KK^*	Th. 3.5	29
$\tilde{u}_n$	Eigenfunction of $\widetilde{K}_N\widetilde{K}_N^*$	Th. 6.5	81
v_n	Eigenfunction of K^*K	Th. 3.5	29
$\tilde{v}_n$	Eigenfunction of $\widetilde{K}_N^*\widetilde{K}_N$	Th. 6.5	81

$\mathcal{V}, \mathcal{W}$	Vector spaces	–	11
$\mathcal{X}$	Hilbert space, typically the domain	Sec. 2.5	18
$\mathcal{Y}$	Hilbert space, typically the range	Sec. 2.5	18
1_I	The characteristic function on I, $t \mapsto 1$ if $t \in I$, 0 otherwise	Eq. (4.13)	46
δ_{ij}	Kronecker delta, $\delta_{ij} = 1$ for $i = j$, 0 otherwise	Eq. (3.9)	30
$\partial_s^n k$	The nth partial derivative of k with respect to s	Sec. 4.5.2	52
λ, λ_n	Eigenvalues	Sec. 3.1	25
μ_n	Singular value	Def. 3.4	29
$\mu_n(K)$	The nth singular value of K (ordered non-decreasingly)	Th. 3.6	31
φ, φ_n	Eigenvector/-function	Sec. 3.1	25
(ϕ_n)	Orthonormal sequence in X	Sec. 6.1	73
(ψ_m)	Orthonormal sequence in Y	Sec. 6.1	73
σ_i^N	The ith singular values of $\widetilde{K}_N$	Th. 6.2	79
$\Pi_{\mathcal{X}}$	Orthogonal projection operator onto $\mathcal{X}$	Eq. (2.11)	18
(f,g)	For f, g in a Hilbert space: The inner product defined	Def. 2.6	15
$(\mathbf{x},\mathbf{y})$	For vectors $\mathbf{x},\mathbf{y}$: $(\mathbf{x},\mathbf{y}) = \mathbf{y}^H\mathbf{x}$	Eq. (2.9)	16
T^*	The adjoint of T	Eq. (2.12)	18
$\mathcal{M}^\perp$	Orthogonal complement	Def. 2.10	17
$\otimes$	Tensor product, $f \otimes g(x,y) = f(x)g(y)$	Eq. (A.2)	135
$\|f\|_p$	The p-norm of f	Eq. (2.2)	13
$\|T\|$	The operator norm ($T \in B(\mathcal{V},\mathcal{W})$)	Eq. (2.3)	14
$\|x\|$	When $x \in \mathcal{V}$ and $\mathcal{V}$ is a normed vector space: The norm defined in $\mathcal{V}$	Eq. (2.1)	12
$\|x\|$	For x in a Hilbert space: The norm induced by the inner product	Eq. (2.6)	15
$\|\mathbf{x}\|_p$	Vector p-norm, $\|\mathbf{x}\|_p^p = \sum_n \lvert\mathbf{x}_n\rvert^p$	Eq. (7.7)	98
$\|\mathbf{A}\|_p$	Matrix p-norm, $\|\mathbf{A}\|_p = \max_{\mathbf{x}\neq\mathbf{0}}\{\|\mathbf{A}\mathbf{x}\|_p/\|\mathbf{x}\|_p\}$	Eq. (7.12)	100

Contents

CHAPTER 1

Introduction

introduction, *the act of introducing, or bringing to notice*
— WEBSTER'S REVISED UNABRIDGED DICTIONARY

Many mathematical problems in all kinds of sciences lead to linear systems of equations, and there exists a lot of methods for solving them numerically. But when efficiency and accuracy is important, the best method to choose is heavily dependent on the problem at hand.

1.1 Linear Inversion

Many physical systems can be described using a linear model. An abstract way to view such a system is illustrated in Figure 1.1. Given some input and a linear model, the output can be calculated. For the

$$\boxed{\text{Input}} \times \boxed{\text{Model}} \rightarrow \boxed{\text{Output}}$$

Figure 1.1: *An abstract way to view an input/output system governed by a model.*

models we will be looking at, this will be quite easy. The situation changes, however, when the *output* is given and either the input or the model is unknown. This is known as a *linear inverse problem*.

To link this to actual real life examples, let us consider some examples:[1]

Geological Prospecting. The general problem of geological prospecting is to determine the location, shape and constitution of subterranean bodies from measurements at the surface of the earth. A concrete example of this could be that a magnetic material is located at some depth below the earth's surface. A model can now be set up that describes how strong the magnetism is at each position at the surface. The inverse problem is now: Given this model and magnetic measurements at the surface, what is the shape of the body underground?

Tomography. Assume an object is illuminated with radiation of known intensity. To some extent, the beam is absorbed inside the object, and the final intensity of the beam can be measured on the far side of the object. If this object is a part of the human body, e.g. the brain, this is known as tomography. Here, the input will be known (the incoming intensity) and the output will be known (the out-coming intensity). The inverse problem is here to determine the model, that is, the object.

Image reconstruction. Suppose a picture of some stars is taken from the earth and digitized. Stars that should be small isolated dots will typically be "smudged" to some extent because of the light's passage through the atmosphere. A model can now be derived that describes how each point is being smudged or blurred. The inverse problem is now clear: Given a blurred digitized image and a model of the blurring, reconstruct the true, unblurred image.

[1]These examples, and a lot of others, can be found in [Gro93].

All of the above problems can be described mathematically by

$$Kf = g,$$

where f represents the input, g the output and K the model. So the challenge here is to "just" calculate $g = K^{-1}f$. But the matter is complicated by two important factors: K typically operates on infinite dimensional spaces and K turns out to be represented by *integral operators*, which have the unfortunate property of being *compact*. These two factors ensure that the inverse operator, K^{-1}, will be *discontinuous*. So small perturbations of f can lead to arbitrarily large perturbations in g. In heat conduction, for instance, this property of the inverse reflects the fact that heat conduction is an *irreversible* physical process (see [Kre99] page 267). These kinds of problems are called *ill-posed* and this is exactly where the difficulties/challenges lie.

Now suppose that such an ill-posed inverse problem is given and that we wish to solve it using a computer. As mentioned, K is most often infinite dimensional and that is impossible to represent on a computer. So we have to *discretize* the problem, that is, project the infinite dimensional problem onto a finite dimensional problem. There are many ways to do this, but they all end up with an innocent looking linear system of equations:

$$\mathbf{A}\mathbf{x} = \mathbf{b}.$$

But the system is not innocent at all and these systems are often termed *discrete ill-posed problems*. Although the problem is finite dimensional and $\mathbf{A}$ may be (mathematically) invertible, it is far from easy to solve. The discontinuity of K^{-1} is reflected in this system by $\mathbf{A}$ being "nearly singular", that is, having a very large condition number. This is quite unfortunate since $\mathbf{b}$ is often affected by noise or measuring errors and computers only have finite floating point precision. When solving the system, the noise and rounding errors can disturb more and more, and the solution can easily end up being completely useless.

This thesis deals with one class of methods to solve such systems in order to arrive at a usable solution. But before we dive into this, let

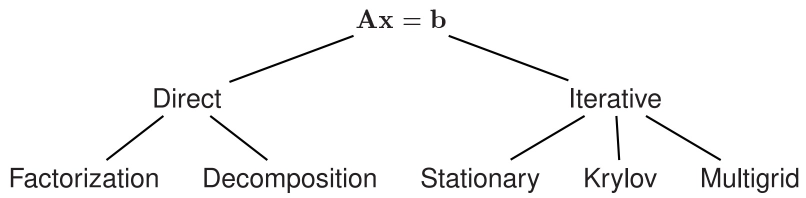


Figure 1.2: *Classification of methods to solve linear systems of equations.*

us put some perspective on things by giving a quick survey on how to solve linear systems of equations in general.

1.2 Solving Linear Systems of Equations, a Quick Survey

The number of methods available to solve a linear system of equations is enormous. Let us for ease of reference have the concrete system $\mathbf{Ax} = \mathbf{b}$ in mind where $\mathbf{A}$ denotes the coefficient matrix, $\mathbf{b}$ the right-hand side and $\mathbf{x}$ the solution.[2] To present classes that each known method fits into is an impossible task. Still, Figure 1.2 tries to do just this, but on a very broad scale.

The topmost division distinguishes between *direct* and *iterative* methods. As the adjective direct indicates, these methods require a direct access to all elements of the coefficient matrix $\mathbf{A}$. The amount of work needed to arrive at the solution is (almost always) known beforehand and the solution is exact if the effect of rounding errors are neglected.

Direct methods can be further divided into *factorization* and *decomposition* methods. Factorization methods factorizes $\mathbf{A}$ into the product of a number of matrices that each make a system easier to solve, given a right-hand side. Such matrices can be triangular, diagonal

[2]Note however, that a linear system of equations need not explicitly be expressed this way.

or orthogonal and the methods include LU-, QR- and Cholesky factorization. Decomposition methods are also a form of factorization, but they simultaneously extract important information relating to e.g. rank, null-space, singular values and so on. The cost is that they typically require more work than factorization methods. Popular methods include Rank Revealing QR, eigenvalue decomposition (diagonalization) and Singular Value Decomposition, SVD.

Instead of direct methods one can turn to iterative methods—the rightmost branch in the tree shown in the figure. These typically start out with a starting guess (possibly the zero vector) and then iterate by repeating identical or similar operations. The hope is now that this process produces a sequence of vectors that converges to the solution. Note that the solutions found by direct methods almost inevitably will be influenced by rounding-errors, and iterative methods may arrive at an equally accurate solution with much less work. However, the amount of work needed is not known beforehand and the produced sequence of vectors may not converge at all.

Stationary methods is one class of iterative methods. They all stem from the idea of splitting the coefficient matrix like $\mathbf{A} = \mathbf{M} - \mathbf{N}$. Each iteration is now expressed like $\mathbf{Mx}^{(k+1)} = \mathbf{Nx}^{(k)} + \mathbf{b}$, where $\mathbf{M}$ is nonsingular and easy to invert. Stationary methods include Jacobi, Gauss–Seidel and SOR.

Another class of iterative methods is Krylov subspace methods. A broad definition of these could be that the only way the (square) coefficient matrix is involved, is via the matrix-vector products $\mathbf{Az}$ and $\mathbf{A}^H\mathbf{z}$ for arbitrary vectors.[3] In a way, these methods are totally opposite the direct methods in that they require no access to the elements of $\mathbf{A}$, only the way that vectors are *mapped* is used. This can be quite useful since $\mathbf{A}$ need not be represented explicitly and efficient subroutines can be tailor-made to perform the multiplications. The name Krylov subspace method comes from the fact that a vector space spanned by $\mathbf{z}, \mathbf{Az}, \ldots, \mathbf{A}^{k-1}\mathbf{z}$ is called a Krylov subspace. Popular Krylov subspace methods include Conjugate Gradients and GMRES.

[3]The notation $\mathbf{A}^H$ means the Hermitian of $\mathbf{A}$, which is equivalent to the conjugate transposed of $\mathbf{A}$. For real matrices, we have $\mathbf{A}^H = \mathbf{A}^T$.

The last class to mention is that of multigrid methods. The last part of the word, "grid", suggests that these methods often work on two- or three-dimensional grids that have some geometrical relation to the physical problem being solved. The first part of the word, "multi", means that these methods work on multiple grid levels, that is, different grid sizes. By using interpolation and restriction, information on one level can be transferred to other levels. This is sometimes very useful and can lead to superior rates of convergence.

Multigrid methods, however, are mostly used together with e.g. stationary methods, and the distinction between different classes is seldom clear. So-called hybrid methods that combine ideas from the different classes are often used.

Important to note is also that some methods only work for problems of a certain kind. This is most often requirements on $\mathbf{A}$ like positive definiteness or symmetry. Other methods again may work in general, but can be tailor-made to work especially well e.g. for sparse or structured coefficient matrices.

1.3 Regularization

The above survey mentioned methods for solving linear systems of equations in general. When discrete ill-posed problems must be solved, special care has to be taken. An integral operator K will always have a smoothing effect, that is, high-frequency components in the input will be mapped over to vectors that contain high-frequency components with *very low* amplitude.[4] This can be shown to be mimicked by the finite dimension approximation $\mathbf{A}$ so the following holds: If $\mathbf{b}$ contains high-frequency components, $\mathbf{x}$ must contain high-frequency components with *very high* amplitude.

If $\mathbf{b}$ is influenced by noise, as it often is, this statement will apply and the solution will be highly influenced by high-frequency components stemming from the noise. This leads to the concept of *regularization* which is a way to arrive at a solution which is useful but not necessarily the true solution. It is typically done by demanding certain things of the solution. This could be that the solution may only consist of low-frequency components or that the solution must be piecewise constant.

[4] Since K is linear and continuous, the input can be split into a sum of components, map them each using K and add them together again. Such a superpositional principle viewpoint can be used.

Not all the methods mentioned in the previous section can be used directly on discrete ill-posed problem. Some can, however, if the problem is altered in a certain way, and others again can be made to have a regularizing effect in themselves. These lastly mentioned methods are for instance QR-factorization (rank-revealing QR), SVD (truncated SVD) and some iterative methods (by using the iteration step as regularization parameter).

The direct methods work very well but need a lot of computing power and memory for large problems. Iterative methods however, are very well suited for large-scale problems and it is thus interesting to investigate their ability to regularize.

1.4 Motivation for this Project

The Krylov subspace method called conjugate gradients (CG) has proved a very useful iterative method, also for discrete ill-posed problems. It has the restriction, though, that it only works for symmetric, positive definite coefficient matrices. So for general $\mathbf{A}$ it can not be used. This can be remedied by using the method on the normal equations instead, but without explicitly forming $\mathbf{A}^H\mathbf{A}$ and $\mathbf{A}^H\mathbf{b}$. This method is often called CGLS. However, matrix–vector multiplications involving both $\mathbf{A}$ and $\mathbf{A}^H$ must be calculated in each iteration.

Another popular Krylov subspace method is called GMRES. This method works for general $\mathbf{A}$ and in each iteration only multiplication with $\mathbf{A}$ is required. The rate of convergence of this method is quite well understood, but only for well-posed problems. This leads us to the question that kicked off this project: "How well does GMRES work for discrete ill-posed problems?" This is largely what the present thesis is all about.

Other interesting aspects quickly became very relevant. Since the method only has powers of **A** to "work with", the eigenvalues of **A** are very important. So to know how these are distributed is essential. Since discrete ill-posed problems most often come from infinite dimensional integral operators, what can we learn from these? At the same time the ill-posedness of the linear equations is due to the fact that (almost all) integral operators are *compact*. This lead to some very important questions:

- Is it possible to generalize the properties of integral operators to a more abstract setting of compact operators?
- Is there a connection between the eigenvalues of a compact operator and those of a matrix approximation? This is quite well understood for singular values and -vectors, but the just posed question is not fully explored.
- Which operators have only a finite number of eigenvalues or none at all?

Note that GMRES should only be used for non-Hermitian **A**, otherwise more efficient methods are available. So most focus will be put on operators that are not assumed to be symmetric (self-adjoint).

As mentioned earlier, it is well understood how well GMRES works for well-posed problems. This type of analysis typically only looks at key properties of **A** and deals with convergence for arbitrary right-hand sides. When dealing with discrete ill-posed problems it turns out that the rate of convergence is very sensitive to the right-hand side. So this is another subject that this thesis will explore.

1.5 Outline

We start off by providing a theoretical foundation. The remaining chapters will use a lot of basic, but important, functional analysis, and most of it will be presented in this second chapter.

Chapter 3 will present the quantities eigenvalues and singular values. These turn out to be very important, not only in this thesis, but also when dealing with ill-posed problems in general.

Chapter 4 looks at compact operators with certain characteristics. The largest part of the chapter will deal with integral operators and the kernels of these. Using the properties of the operators and kernels, they are investigated with especially the behavior of the eigenvalues and singular values in focus.

Chapter 5 will give some precise definitions of ill-posedness and regularization. This will make it clearer what the problem/challenge is and what we are aiming for. The chapter is deliberately postponed to here because some of the theory of the previous chapters are needed.

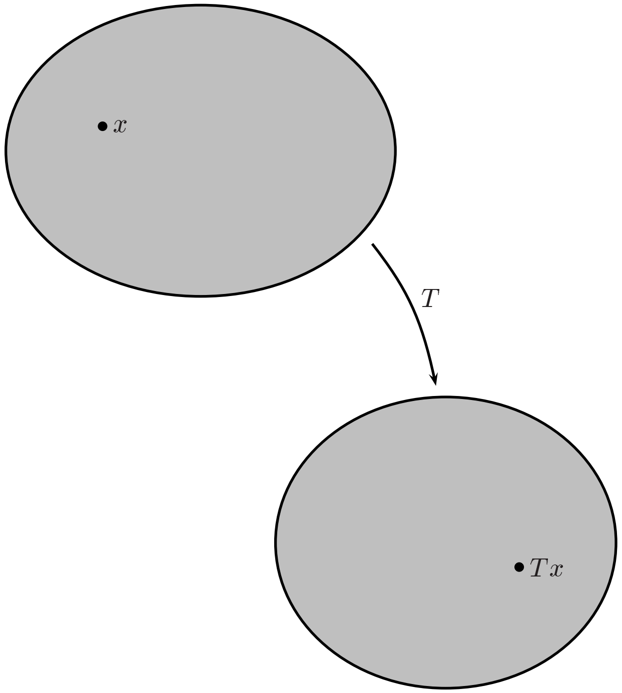
Chapter 6 will address discretization, or more specifically, Galerkin discretization. To computationally deal with these problems, they have to be made finite dimensional. But how well do these finite dimensional operators approximate the original operator? Special focus will be put on the eigenvalues and singular vectors (and corresponding vectors).

Chapter 7 turns to look at Krylov subspaces and a class of methods, whose solutions lie in these subspaces. A theoretical investigation will be made in order to find out when solutions can be found at all, the maximal number of iterations and related subjects. In particular the method GMRES will be looked upon concerning convergence analysis and finally, related methods such as MINRES and CG will be mentioned.

Chapter 8 looks at how to actually implement GMRES. By reformulating the original problem, a quite simple and efficient algorithm can be produced. A number of related algorithms and implementation issues will also be discussed.

Chapter 9 finally attempts to collect all the threads and summarize the most important results. A list of subjects for further study will be included.

The appendix contains a number of sections. Most are referenced to from appropriate places in the thesis, but one is of special importance, namely the "Examples" section, which provides a number of concrete examples that illustrate many important aspects of the theory presented throughout this thesis.



CHAPTER 2

Theory of Linear Operators

theory, *a scheme of the relations subsisting between the parts of a systematic whole*
— WEBSTER'S REVISED UNABRIDGED DICTIONARY

operator, *one who, or that which, operates or produces an effect*
— WEBSTER'S REVISED UNABRIDGED DICTIONARY

This chapter will provide the theoretic foundation for the rest of the thesis. Most of the contents is basic functional analysis, and very few proofs will be provided (see e.g. [Ped00] or [Kre99] for proofs and additional information). At the same time, much of the used notation will also be introduced here.

2.1 Metric Spaces

One of the goals in this section is to continuously apply structure to a vector space until it has all properties necessary for our later use.

We will start of by requiring that a vector space $\mathcal{V}$ is equipped with a *metric*. Such a vector space is called a *metric space*. A metric is a mapping $d : \mathcal{V} \times \mathcal{V} \mapsto [0, \infty[$ that fulfills symmetry, the triangle inequality and $d(x, y) = 0 \Leftrightarrow x = y$ for all element pairs.

A set $M \subset \mathcal{V}$ is termed *open* if for every $x_0 \in M$ there exists an $r > 0$ such that $\{x \in \mathcal{V} \mid d(x_0, x) < r\} \subset M$. A set is defined to be *closed* if its complement is open. The *closure* $\overline{M}$ of a set M is the smallest closed set containing M.

Convergence of a sequence $(x_n) \subset \mathcal{V}$ is definable using the metric. If $d(x_n, x) \to 0$ for some $x \in \mathcal{V}$ when $n \to \infty$ we say that (x_n) *converges* to x and write $x_n \to x$ for $n \to \infty$. A looser notion of a sequence possibly approaching a limit is that of a *Cauchy sequence*, which just requires that $d(x_n, x_m) \to 0$ for $n, m \to \infty$. When it happens that every Cauchy sequence is actually convergent, the space is called *complete*.

Another important concept is denseness. Let $\mathcal{V}$ be a metric space with metric d and let $\mathcal{A}, \mathcal{B} \subset \mathcal{V}$. If for all $b \in \mathcal{B}$ and $\epsilon > 0$ there exists an $a \in \mathcal{A}$ such that $d(a, b) < \epsilon$, we say that $\mathcal{A}$ is *dense* in $\mathcal{B}$.

Assume two metric spaces $\mathcal{V}$ and $\mathcal{W}$ are given with metrics $d_{\mathcal{V}}$ and $d_{\mathcal{W}}$ respectively. If there exists a bijective map $T : \mathcal{V} \mapsto \mathcal{W}$ such that $d_{\mathcal{W}}(Tx, Ty) = d_{\mathcal{V}}(x, y)$ for all $x, y \in \mathcal{V}$ we say that $\mathcal{V}$ and $\mathcal{W}$ are *isometric*. This finally leads us to the notion of completion:

Theorem 2.1 *Let $\mathcal{V}$ be a metric space with metric $d_{\mathcal{V}}$. Then there is a complete metric space $\mathcal{W}$ with metric $d_{\mathcal{W}}$ and a dense set $\widetilde{\mathcal{W}} \subset \mathcal{W}$ such that $\mathcal{V}$ and $\widetilde{\mathcal{W}}$ are isometric.*

The space $\mathcal{W}$ is denoted the ***completion*** *of $\mathcal{V}$.*

2.2 Normed Vector Spaces

A normed space is a vector space $\mathcal{V}$ equipped with a norm. A norm is a mapping $\|\cdot\| : \mathcal{V} \mapsto [0, \infty[$ which fulfills

$$\|x + y\| \leq \|x\| + \|y\| \tag{2.1a}$$
$$\|\alpha x\| = |\alpha|\|x\| \tag{2.1b}$$
$$\|x\| = 0 \Rightarrow x = 0 \tag{2.1c}$$

for all $x, y \in \mathcal{V}$ and $\alpha \in \mathbb{C}$.

Given a normed vector space, a metric can be induced in a natural way simply by defining $d(x, y) = \|x - y\|$. We can now have a more structured vector space:

Definition 2.2 *A normed vector space that is complete in the metric induced by the norm is called a* ***Banach space****.*

The *support* of a function $f : \Omega \mapsto \mathbb{C}$, $\Omega \subset \mathbb{R}^k$, is the closure of the set $\{x \in \Omega \mid f(x) \neq 0\}$. The space $C_0(\Omega)$ is the space of continuous functions with bounded support. This space can be equipped with the so-called p-norms. These are maps $\|\cdot\|_p : C_0(\Omega) \mapsto [0, \infty[$, $p \geq 1$, given by

$$\|f\|_p = \left(\int_\Omega |f(x)|^p dx \right)^{\frac{1}{p}}. \tag{2.2}$$

It may not be obvious that these in fact are norms, but it can be proved straightforwardly. We can now define some very important vector spaces:

Definition 2.3 *The vector space $L^p(\Omega)$, $\Omega \subset \mathbb{R}^k$ where $p \geq 1$, is the completion of $C_0(\Omega)$ in the metric induced by the p-norm.*

Especially L^2 will be used extensively throughout this thesis. In connection with this function space and the norms just introduced, we have the following important inequality.

Theorem 2.4 (Cauchy-Schwartz' inequality) *For $f, g \in L^2(\Omega)$ we have*

$$\|fg\|_1 \leq \|f\|_2 \|g\|_2.$$

Spaces of sequences can be defined in the following way:

Definition 2.5 *The space of sequences (real or complex) l^p, $p \geq 1$, consists of sequences (x_n) satisfying*

$$\sum_{n=1}^{\infty} |x_n|^p < \infty.$$

The norm in l^p is

$$\|(x_n)\|_p = \left(\sum_{n=1}^{\infty} |x_n|^p\right)^{\frac{1}{p}}.$$

As we shall later see, there is a tight connection between L^2 and l^2.

2.3 Bounded Linear Operators

An operator in its most abstract form is a mapping between two normed spaces. Consider an operator $T : \mathcal{V} \mapsto \mathcal{W}$. Important properties that T can have are:

- T is *linear* if
$$\forall x, y \in \mathcal{V}\ \forall \alpha, \beta \in \mathbb{C} \quad T(\alpha x + \beta y) = \alpha T(x) + \beta T(y).$$
- T is *continuous at a point $x \in \mathcal{V}$* if
$$\forall \epsilon > 0\ \exists \delta > 0\ \forall y \in \mathcal{V} \quad \|x - y\| < \delta \Rightarrow \|Tx - Ty\| < \epsilon.$$
- T is *continuous* if T is continuous at all points $x \in \mathcal{V}$.
- T is *bounded* if
$$\exists M > 0\ \forall x \in \mathcal{V} \quad \|Tx\| \leq M\|x\|.$$

When an operator is linear, the three last properties are actually *equivalent*.

The set of linear and bounded operators from $\mathcal{V}$ into $\mathcal{W}$ will be denoted $B(\mathcal{V}, \mathcal{W})$ (when $\mathcal{V} = \mathcal{W}$ we write $B(\mathcal{V})$). It is a vector space and can be equipped with a norm, the *operator norm*:

$$\|T\| = \sup\left\{\|Tx\| \mid \|x\| \leq 1\right\}. \tag{2.3}$$

Note that $\|x\|$ refers to the norm defined in $\mathcal{V}$ and $\|Tx\|$ refers to the norm defined in $\mathcal{W}$.

When $\mathcal{V}$ is a normed space and $\mathcal{W}$ is a Banach space, then $B(\mathcal{V}, \mathcal{W})$ is itself a Banach space.

2.4 Hilbert Spaces

We continue to apply structure by introducing inner products. This enables us, among other things, to define when vectors are orthogonal.

Definition 2.6 *Let $\mathcal{V}$ be a vector space. An **inner product** is a mapping $(\cdot,\cdot) : \mathcal{V} \times \mathcal{V} \mapsto \mathbb{C}$ satisfying*

$$\begin{aligned}
(x, y) &= \overline{(y, x)} && (2.4a)\\
(\alpha x_1 + \beta x_2, y) &= \alpha(x_1, y) + \beta(x_2, y) && (2.4b)\\
(x, y) &\geq 0\,,\ (x, x) = 0 \Leftrightarrow x = 0 && (2.4c)
\end{aligned}$$

for all $x_1, x_2, x, y \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{C}$.

In $L^2(\mathbb{R})$ we have the inner product

$$(f, g) = \int_{\mathbb{R}} f(x)\overline{g(x)}dx. \tag{2.5}$$

An inner product also gives rise to an *induced norm*, defined by

$$\|x\| = (x, x)^{\frac{1}{2}} \tag{2.6}$$

which is well-defined for any vector space equipped with a norm. For $L^2(\mathbb{R})$ the induced norm becomes

$$\|f\| = (f, f)^{\frac{1}{2}} = \left(\int_{\mathbb{R}} f(x)\overline{f(x)}dx\right)^{\frac{1}{2}} = \left(\int_{\mathbb{R}} |f(x)|^2 dx\right)^{\frac{1}{2}} \tag{2.7}$$

which is exactly the 2-norm defined for $C_0(\mathbb{R})$ functions in Equation (2.2).

The ground has now been laid to define a Hilbert space:

Definition 2.7 *A vector space, with an inner product, that is a Banach space with respect to the induced norm is called a **Hilbert space**.*

The Banach space $L^2(\mathbb{R})$ is also a Hilbert space. It has an inner product defined by (2.5), which in turn defines a norm, which in turn defines a metric. And by *construction*, $L^2(\mathbb{R})$ is a Banach space since every element of $L^2(\mathbb{R})$ is the limit function of a Cauchy sequence consisting of $C_0(\mathbb{R})$ functions. By introducing inner products similar to that in Equation (2.5), the spaces $L^2(\Omega)$ where $\Omega \subset \mathbb{R}^n$ can also be shown to be Hilbert spaces.

For l^2, the space of square summable sequences, the inner product is defined as

$$((x_n),(y_n)) = \sum_{k=1}^{\infty} x_k \overline{y_k}. \tag{2.8}$$

For finite dimensional (coordinate) vectors, this sum becomes finite and is similar to the well known inner product in linear algebra,

$$(\mathbf{x},\mathbf{y}) = \sum_{k=1}^{n} \mathbf{x}_k \overline{\mathbf{y}_k} = \mathbf{y}^H \mathbf{x}, \quad \text{for } \mathbf{x},\mathbf{y} \in \mathbb{C}^n. \tag{2.9}$$

From linear algebra it is also well known that every vector of a given (finite dimensional) vector space can be uniquely identified in terms of its coordinates with respect to some orthonormal basis. This can be extended to general Hilbert spaces by the following theorem.

Theorem 2.8 *Let $\mathcal{H}$ be a Hilbert space. Then $\mathcal{H}$ is isometric isomorphic to l^2.*

Two normed vector spaces are isometric isomorphic when there exists bijective, linear and isometric operators mapping one to the other.

So considering abstract vectors from a Hilbert space or their coordinates with respect to some basis is "the same". This is furthermore supported by the following proposition.

Proposition 2.9 *An orthonormal sequence (x_k) in a Hilbert space $\mathcal{H}$ is an*

orthonormal basis if and only if one of the following conditions hold:[1]

$$(x,y) = \sum_{k=1}^{\infty}(x,x_k)(x_k,y) \quad \textit{for all } x,y \in \mathcal{H} \tag{2.10a}$$

$$\|x\|^2 = \sum_{k=1}^{\infty} |(x,x_k)|^2 \quad \textit{for all } x \in \mathcal{H} \tag{2.10b}$$

$$\forall k\ (x,x_k) = 0 \quad \Leftrightarrow \quad x = 0. \tag{2.10c}$$

Apart from the fact that an inner product can induce a norm, it can also indicate when vectors are orthogonal: The (non-zero) vectors $x, y \in \mathcal{H}$ are orthogonal exactly when $(x,y) = 0$. The notion of orthogonality can be extended to subsets in the following way.

Definition 2.10 *Let $\mathcal{M}$ and $\mathcal{N}$ be nonempty subsets of a Hilbert space $\mathcal{H}$. We say that $\mathcal{M}$ and $\mathcal{N}$ are **orthogonal** and write $\mathcal{M} \perp \mathcal{N}$ if $(x,y) = 0$ for all $x \in \mathcal{M}$ and $y \in \mathcal{N}$.*

*For a nonempty subset $\mathcal{M}$ of $\mathcal{H}$, we define the **orthogonal complement** to $\mathcal{M}$, denoted $\mathcal{M}^\perp$, by*

$$\mathcal{M}^\perp = \{y \in \mathcal{H} \mid (x,y) = 0 \textit{ for all } x \in \mathcal{M}\}.$$

Notice that since an inner product is linear in the first variable, $\mathcal{M}^\perp$ is obviously closed.

Let $\mathcal{M}$ and $\mathcal{N}$ be closed subspaces of a Hilbert space $\mathcal{H}$ and $\mathcal{M} \perp \mathcal{N}$. We define the *orthogonal sum* of $\mathcal{M}$ and $\mathcal{N}$, denoted $\mathcal{M} \oplus \mathcal{N}$, by

$$\mathcal{M} \oplus \mathcal{N} = \{z \in \mathcal{H} \mid z = x + y, x \in \mathcal{M}, y \in \mathcal{N}\}.$$

The next theorem states that every Hilbert space can be seperated into two subsets orthogonal to each other: A closed subset and "the rest".

Theorem 2.11 (Projection Theorem) *If $\mathcal{M}$ is a closed subspace of a Hilbert space $\mathcal{H}$, then $\mathcal{H} = \mathcal{M} \oplus \mathcal{M}^\perp$.*

[1] Equation (2.10b) is known as Parsevals equation.

We also introduce the concept of a *projection operator*. An operator $\Pi \in B(\mathcal{X})$ is a projection operator if and only if $\Pi^2 = \Pi$. Let $\mathcal{X}$ be spanned by the orthonormal basis (ϕ_n). We can now project *orthogonally* onto the space $\mathcal{X}_N = \text{span}\{\phi_1, \phi_2, \ldots, \phi_N\}$ with the following projection operator

$$\Pi_{\mathcal{X}_N} x = \sum_{n=1}^{N} (x, \phi_n)\phi_n. \tag{2.11}$$

Since $\Pi_{\mathcal{X}_N}\phi_1 = \phi_1$ it is easy to see that $\|\Pi_{\mathcal{X}_N}\| = 1$.

2.5 Operators on Hilbert Spaces

We will now look at bounded linear operators from $\mathcal{X}$ into $\mathcal{Y}$, where $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces. First, the *adjoint* operator must be defined.

Theorem 2.12 *Let $T \in B(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces. Then there is a unique operator $T^* \in B(\mathcal{Y}, \mathcal{X})$ satisfying*

$$(Tx, y) = (x, T^*y) \quad \text{for all } x, y \in \mathcal{X} \times \mathcal{Y},$$

and we have $\|T\| = \|T^*\|$. The operator T^* is called the ***adjoint*** *of T.*

The adjoint is a generalization of the Hermitian (or conjugate transpose) of a matrix. When an operator T is represented by a finite matrix $\mathbf{A}$, its adjoint T^* is represented by $\mathbf{A}^H$. This will be proved more formally later.

An operator T can have two very important properties relating to its adjoint. If $T = T^*$ we call T *self-adjoint* and if $T^*T = TT^*$ we call T *normal*. Note that self-adjoint operators are also normal.

Consider now an operator $T \in B(\mathcal{X}, \mathcal{Y})$ where $\mathcal{X}$ and $\mathcal{Y}$ are Hilbert spaces. We now wish to use the projection theorem to split up $\mathcal{Y}$ into the range, $\mathcal{R}(T)$, and the rest. But $\mathcal{R}(T)$ is not necessarily closed, so the best we can do is to write

$$\mathcal{Y} = \overline{\mathcal{R}(T)} \oplus \mathcal{R}(T)^{\perp}.$$

This can be rewritten. Let $y \in \mathcal{R}(T)^{\perp}$, which means that $(Tx, y) = 0$ for all $x \in \mathcal{X}$. By using the definition of the adjoint, this is equivalent to $(x, T^*y) = 0$ for all $x \in \mathcal{X}$. But this means that $T^*y = 0$ so $y \in \mathcal{N}(T^*)$. Now have this formulation

$$\mathcal{Y} = \overline{\mathcal{R}(T)} \oplus \mathcal{N}(T^*). \tag{2.12}$$

Before defining one of the most important concepts of this thesis, compact operators, we must define what a compact *set* is.

Definition 2.13 *A set $\mathcal{S}$ in a normed space $\mathcal{V}$ is called* ***compact*** *if every sequence in $\mathcal{S}$ has a subsequence that converges to an element of $\mathcal{S}$.*

A compact set is closed and bounded and in finite dimensions we furthermore have that a closed and bounded set is compact, i.e. they are equivalent. We can now define a compact operator:

Definition 2.14 *Let $\mathcal{V}$ and $\mathcal{W}$ be normed spaces. An operator $K \in B(\mathcal{V}, \mathcal{W})$ is termed a* ***compact operator*** *if $\overline{K(\mathcal{A})}$ is compact in $\mathcal{W}$ for all bounded sets $\mathcal{A} \subset \mathcal{V}$.*

If the dimension of $K(V)$ is finite we say that K has finite rank. Consider such an operator and a bounded subset $\mathcal{A} \subset \mathcal{V}$. The set $\overline{K(\mathcal{A})}$ is closed by construction and it is also bounded since K is bounded. Hence, $\overline{K(\mathcal{A})}$ is compact and so is the operator K. In other words: Every bounded linear operator of finite rank is compact.

We now state the following important theorem.

Theorem 2.15 *Let $\mathcal{X}$, $\mathcal{Y}$ be Hilbert spaces and assume that (K_n) is a sequence of compact operators in $B(\mathcal{X}, \mathcal{Y})$ converging to an operator K (in the operator norm). Then K is compact.*

This theorem shows that the set of compact operators in $B(\mathcal{X}, \mathcal{Y})$ is closed. It also follows that if a given operator K can be approximated arbitrarily well using operators of finite rank, it must be compact.

One can show that an operator is compact if and only if its adjoint is compact. Furthermore, any product, both from left and right, of a compact operator and a bounded linear operator is compact (see pages 485–486 in [DS64]).

We now introduce the concept of weak convergence, which turns out to have special special properties for compact operators.

A sequence $(x_n) \subset \mathcal{H}$ is said to converge *weakly* to x if $(x_n, y) \to (x, y)$ for all $y \in \mathcal{H}$. The choice of the adjective *weak* is not accidental. Assume that $x_n \to x$, that is, $\|x_n - x\| \to 0$. Then we have from Cauchy-Schwartz' inequality

$$|(x_n, y) - (x, y)| = |(x_n - x, y)| \leq \|x_n - x\|\|y\| \to 0,$$

so (usual) convergence implies weak convergence.

We can now prove the following theorem.

Theorem 2.16 *Let $\mathcal{H}$ be a Hilbert space and let (x_n) be a weakly convergent sequence with (weak) limit x. If $K \in B(\mathcal{H})$ is compact then (Kx_n) converges in norm to Kx.*

Proof: The assumption of weak convergence implies for all $y \in \mathcal{H}$ that

$$(Kx_n, y) = (x_n, K^*y) \to (x, K^*y) = (Kx, y) \quad \text{for } n \to \infty,$$

so (Kx_n) converges weakly to Kx. Since usual convergence implies weak convergence, Kx is the only possible limit. So assume that (Kx_n) does not converge to Kx. Then it is possible to extract a subsequence (Kx_{n_k}) of (Kx_n) such that

$$\|Kx_{n_k} - Kx\| > \delta \quad \text{for all } k \in \mathbb{N}$$

and some $\delta > 0$. But since (x_n) is bounded and K is compact, we can find a subsequence $(Kx_{n_{k_l}})$ of (Kx_{n_k}) that is (usually) convergent to a $y \in \mathcal{H}$, but since $(Kx_{n_{k_l}})$ converges weakly to Kx we must have that $y = Kx$. This is not possible according to the inequality above, hence $Kx_n \to Kx$. □

Note that in a Hilbert space $\mathcal{H}$ with arbitrary orthonormal basis (e_n) we have $x = \sum_n (x, e_n)e_n$ for all $x \in \mathcal{H}$. This means that e_n converges weakly to 0 as $n \to \infty$. From the theorem above, this implies that $Ke_n \to 0$ for compact K.

The following theorem provides a result that shows exactly when an operator is bounded and when it is compact.

Theorem 2.17 *Let $T : \mathcal{X} \mapsto \mathcal{Y}$ be an operator with orthonormal sequence $(v_n) \subset \mathcal{X}$ and orthonormal basis $(u_n) \subset \mathcal{Y}$. Let furthermore (μ_n) be a sequence of complex numbers. Define T by*

$$Tx = \sum_{n=1}^{\infty} \mu_n (x, v_n) u_n.$$

Then T is bounded if and only if (μ_n) is bounded, and T is compact if and only if $\mu_n \to 0$.

Proof: Assume that (μ_n) is bounded. That means that a C exists such that $\mu_n \leq C$ for all n. Then from Parsevals equality, Equation (2.10b), we get

$$\|Tx\|^2 = \sum_{n=1}^{\infty} |\mu_n|^2 |(x, v_n)|^2 \leq C^2 \sum_{n=1}^{\infty} |(x, v_n)|^2 \leq C^2 \|x\|^2$$

which means that T is bounded. On the other hand, if T is bounded, a C exists such that $\|Tx\| \leq C\|x\|$ for all $x \in \mathcal{X}$ and from the inequality above, we see that (μ_n) must then be bounded.

Assume now that $\mu_n \to 0$ and define the sequence of operators (T_k) by

$$T_k x = \sum_{n=1}^{k} \mu_n (x, v_n) u_n.$$

Since T_k has finite rank for every k, all T_k's are compact, and since

$$\|Tx - T_k x\|^2 = \sum_{n=k+1}^{\infty} |\mu_n|^2 |(x, v_n)|^2 \leq C_k^2 \|x\|^2$$

where $C_k = sup_{n>k}\{|\mu_n|\}$ we see that

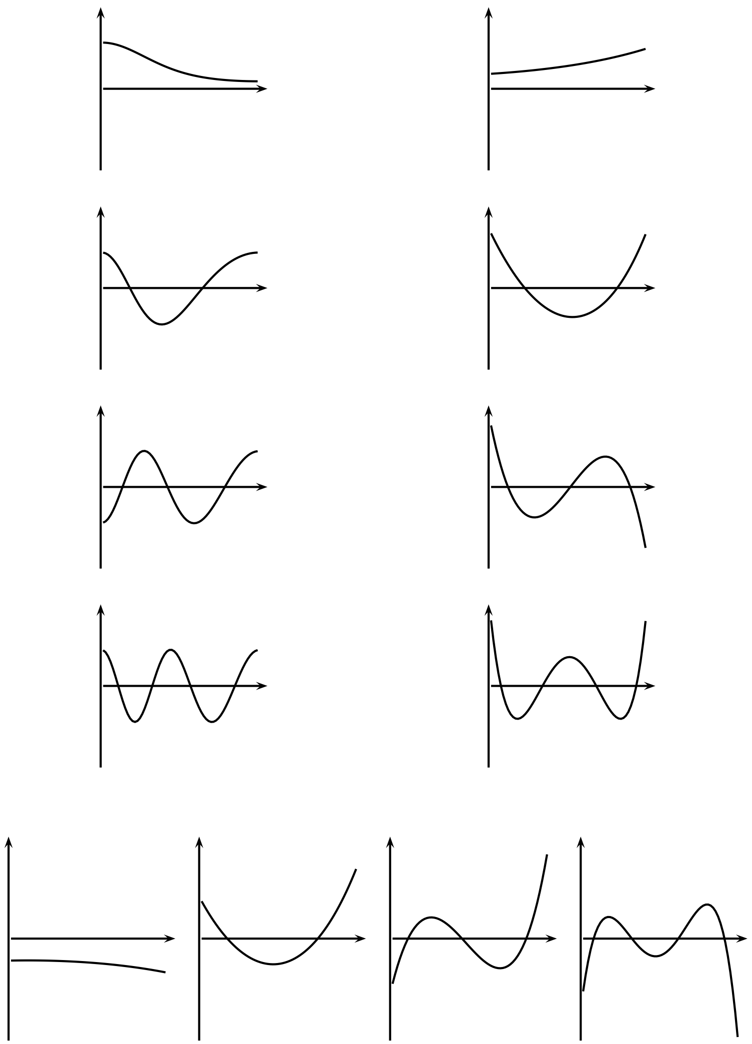
$$\|T - T_k\| \leq C_k \to 0,$$

implying that $T_k \to T$, so T is compact.

Assume instead that $\mu_n \nrightarrow 0$. Then there is a subsequence (μ_{n_k}) of (μ_n) and an $\epsilon > 0$ such that $|\mu_{n_k}| > \epsilon$ for all k. Consider the corresponding subsequences (v_{n_k}) of (v_n) and (u_{n_k}) of (u_n). They are orthonormal and since

$$\|Tv_{n_i} - Tv_{n_j}\|^2 = \|\mu_{n_i}u_{n_i} - \mu_{n_j}u_{n_j}\|^2 = |\mu_{n_i}|^2 + |\mu_{n_j}|^2 > 2\epsilon^2$$

for all $i \neq j$, (Tv_n) can have no subsequences that are Cauchy, hence (Tv_{n_k}) does not converge to 0, and T is not compact (cf. the remark following Theorem 2.16). □



CHAPTER 3

Eigenvalues and Singular Values

singular, *being the only one of a kind; unique*
— THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE

This chapter contains a general view of eigenvalues and singular values for compact linear operators. We will look at how these values behave for concrete classes of operators in the next chapter.

3.1 Eigenvalues

Given a linear operator $T : \mathcal{H} \mapsto \mathcal{H}$, then if

$$Tx = \lambda x$$

has non-trivial solutions for some $\lambda \in \mathbb{C}$ we call λ an eigenvalue of T. The corresponding x's that satisfy the equation, are called eigenvectors. Note that, because of the linearity of T, the eigenvectors will form a vector space called the eigenspace corresponding to λ.

3.1.1 The Spectrum

The set of eigenvalues is a subset of the *spectrum* of T. The spectrum can be introduced by defining the *resolvent* of T,

$$R_\lambda(T) = (T - \lambda I)^{-1} \tag{3.1}$$

which is defined for those $\lambda \in \mathbb{C}$ for which $(T - \lambda I)$ is injective. We now have the following definition.

Definition 3.1 *The* ***resolvent set*** *for T, $\rho(T)$, is the set of $\lambda \in \mathbb{C}$ for which $R_\lambda(T)$ exists as a densely defined and bounded operator on $\mathcal{H}$. The complement $\sigma(T) = \mathbb{C} \setminus \rho(T)$ is called the* ***spectrum*** *for T.*

The spectrum can be divided into three disjoint subsets. If $R_\lambda(T)$ does not exist it means that $(T - \lambda I)$ is not injective. This implies that we have $(T - \lambda I)x = (T - \lambda I)y$ for some $x, y \in \mathcal{H}$ were $x \neq y$. So $(T - \lambda I)(x - y) = 0$ which means that λ is an eigenvalue with corresponding eigenvector $x - y$ ($\neq 0$). The subset of $\sigma(T)$ consisting of all eigenvalues is called the *point spectrum* of T.

The set of λ's for which $R_\lambda(T)$ exists and is densely defined but unbounded, is called the *continuous spectrum* for T.

Finally, the set of λ's for which $R_\lambda(T)$ exists but is not densely defined is called the *residual spectrum* for T.

3.1.2 Eigenvalues of Compact Operators

If a compact operator $K \in B(\mathcal{X})$ is defined on a finite N-dimensional space, we know from linear algebra that the spectrum of K consist of exactly N eigenvalues (counting with multiplicity). The interesting case is when $\mathcal{X}$ is infinite dimensional:

Theorem 3.2 *Let $K \in B(\mathcal{X})$ be a compact operator on an infinite dimensional normed space $\mathcal{X}$. Then $\lambda = 0$ belongs to the spectrum $\sigma(K)$ and $\sigma(K) \setminus \{0\}$ consists of at most a countable set of eigenvalues with no point of accumulation except, possibly, $\lambda = 0$.*

Furthermore, the eigenspace of every non-zero eigenvalues will be finite dimensional.

Proof: See Theorem 3.9 in [Kre99]. □

Note that if K is defined on an infinite dimensional space, $\lambda = 0$ will *always* belong to the spectrum. It can belong to any part of the spectrum though, and if $\lambda = 0$ is an eigenvalue, the corresponding eigenspace may be infinite dimensional.

We now turn to the case of a compact and self-adjoint operator. Such an operator has some especially nice properties, very similar to those of a symmetric matrix in linear algebra. The following spectral theorem summarizes these properties.

Theorem 3.3 (Spectral Theorem for Compact Self-Adjoint Operators) *Let T be a compact and self-adjoint operator on a Hilbert space $\mathcal{H}$. Then $\mathcal{H}$ has an orthonormal basis (e_n) consisting of eigenvectors for T. If $\mathcal{H}$ is infinite dimensional, the corresponding eigenvalues, different from 0, (λ_n) are real and can be arranged in a non-increasing sequence $|\lambda_1| \geq |\lambda_2| \geq \cdots$ where $\lambda_n \to 0$ for $n \to \infty$. Every vector x can be written as*

$$x = \sum_{n \in \mathbb{L}} (x, \varphi_n)\varphi_n + Qx,$$

where $Q : \mathcal{H} \mapsto \mathcal{N}(T)$ projects onto the null-space of T and in this case the mapping becomes

$$Tx = \sum_{n \in \mathbb{L}} \lambda_n (x, \varphi_n)\varphi_n.$$

If the range of T has dimension N we have $\mathbb{L} = \{1, 2, \ldots, N\}$. Is the range infinite dimensional we have $\mathbb{L} = \mathbb{N}$.

Although the proof has been omitted, some aspects of it are important to mention, since it tells us how to pick out the eigenvalues one by one, in non-increasing order. It relies on the fact that for a compact and self-adjoint operator T, defined on a Hilbert space $\mathcal{H}$, we always have (see e.g. [Ped00])

$$|\lambda_1| = \|T\| = \max\{|(T\varphi, \varphi)| \mid \varphi \in \mathcal{H}, \|\varphi\| = 1\}$$

where λ_1 is an eigenvalue of T and obviously the largest one in magnitude. Let a corresponding eigenfunction be denoted φ_1. By constructing $\mathcal{Q}_1 = \{\varphi_1\}^\perp$ we find that T can now be considered as an operator on this new Hilbert space $\mathcal{Q}_1$. And so there must exist a largest eigenvalue in this space. In general:

$$\begin{aligned} |\lambda_n| = \|T\| = \max\{|(T\varphi,\varphi)| \mid \varphi \in \mathcal{Q}_{n-1}, \|\varphi\| = 1\}, \\ \mathcal{Q}_n = \operatorname{span}\{\varphi_1,\varphi_2,\ldots,\varphi_n\}^\perp \end{aligned} \tag{3.2}$$

If $\mathcal{H}$ is finite dimensional this process will stop when $\mathcal{H}$ has been spanned by the eigenvectors—otherwise it will continue infinitely while $|\lambda_n| \to 0$.

We note that when an operator is *normal*, $T^*T = TT^*$, the properties in the spectral theorem for compact self-adjoint operators still hold except that the eigenvalues can now also be complex (Corollary X.4.5 in [DS63]).

When an operator T is non-negative ($(Tx,x) \geq 0$ for all $x \in \mathcal{H}$), it is self-adjoint and can only have non-negative eigenvalues. In this case, a minimax principle can be utilized to obtain a formulation equivalent to the expression (3.2):

$$\lambda_1 = \sup_{\|\varphi\|=1} (T\varphi,\varphi) \tag{3.3a}$$

$$\lambda_{n+1} = \inf_{z_1,\ldots,z_n\in\mathcal{H}} \sup_{\substack{\varphi\perp z_1,\ldots,z_n \\ \|\varphi\|=1}} (T\varphi,\varphi) \tag{3.3b}$$

This result is due to Weyl and Courant (see [Kre99, p. 276] for a proof).

3.2 Singular Values

Singular values have proved a very useful tool in analyzing compact operators. They rely on the simple fact that the operator K^*K is self-adjoint and non-negative for every compact K.

Definition 3.4 *Let $\mathcal{X}$ and $\mathcal{Y}$ be Hilbert spaces and $K \in B(\mathcal{X},\mathcal{Y})$ a compact operator. The (non-negative) square roots of the eigenvalues of $K^*K : \mathcal{X} \to \mathcal{X}$, (μ_n), are called the **singular values** of K.*

The null-spaces of K and K^*K are identical. If $x \in \mathcal{N}(K)$ then clearly $x \in \mathcal{N}(K^*K)$. Assume $x \in \mathcal{N}(K^*K)$. Then $0 = (K^*Kx,x) = (Kx,Kx) = \|Kx\|^2 \Rightarrow Kx = 0 \Leftrightarrow x \in \mathcal{N}(K)$. From this follows that

$$\mathcal{N}(K) = \mathcal{N}(K^*K). \tag{3.4}$$

More insight can be obtained by introducing singular vectors. They have some very nice properties, similar to those from the spectral theorem for self-adjoint operators.

Theorem 3.5 *Let μ_n, $n \in \mathbb{L}$, denote the sequence of non-zero singular values of the compact operator $K \in B(\mathcal{X},\mathcal{Y})$ repeated according to their multiplicity. Then there exists orthonormal sequences $(v_n) \subset \mathcal{X}$ and $(u_n) \subset \mathcal{Y}$ such that*

$$Kv_n = \mu_n u_n \tag{3.5}$$

$$K^*u_n = \mu_n v_n \tag{3.6}$$

for all $n \in \mathbb{L}$. For each $f \in \mathcal{X}$ we have

$$f = \sum_{n\in\mathbb{L}} (f,v_n)v_n + Qf \tag{3.7}$$

where $Q : \mathcal{X} \mapsto \mathcal{N}(K)$ projects onto the null-space of K and the mapping can be written as

$$Kf = \sum_{n\in\mathbb{L}} \mu_n (f,v_n)u_n. \tag{3.8}$$

Proof: Let $\mathbb{L}$ denote the set of indices of all non-zero singular values of K. Let (v_n) be an orthonormal sequence consisting of eigenvectors for K^*K so that $K^*Kv_n = \mu_n^2 v_n$ for all $n \in \mathbb{L}$. Define a second sequence by

$$u_n = \frac{1}{\mu_n} K v_n, \quad n \in \mathbb{L}.$$

This sequence is orthonormal since

$$(u_i, u_j) = \left(\frac{1}{\mu_i}Kv_i, \frac{1}{\mu_j}Kv_j\right) = \frac{1}{\mu_i\mu_j}(K^*Kv_i, v_j) = \frac{\mu_i}{\mu_j}(v_i, v_j) = \delta_{ij}, \tag{3.9}$$

and (3.5) and (3.6) are obviously fulfilled. To prove (3.8) we observe that an arbitrary $f \in \mathcal{X}$ can be written

$$f = \sum_{n\in\mathbb{L}}(f, v_n)v_n + Qf,$$

where $Q : \mathcal{X} \mapsto \mathcal{N}(K^*K) = \mathcal{N}(K)$ projects onto the null-space of K (see (3.4) for the null-space equality). From this follows

$$Kf = \sum_{n\in\mathbb{L}}(f, v_n)Kv_n + KQf = \sum_{n\in\mathbb{L}}\mu_n(f, v_n)u_n.$$

□

Equation (3.8) is often termed the Singular Value Expansion (or Decomposition).

Note that the set $\mathbb{L}$ was introduced for simplicity. Later on, summations of the form $\sum_{n=1}^{\infty}$ will be used. When only a finite number of non-zero singular values exists, the sum should be understood as finite.

Notice the relation to Theorem 2.17 (page 21) which states exactly when an operator, defined as in Equation (3.8), is bounded and when it is compact. The above theorem has now shown that whenever an operator is compact, the mapping *can* be expressed like in (3.8).

Since K^*K is an obvious non-negative operator, the expressions in (3.3) for the eigenvalues of such an operator can appropriately be transferred to this setting. Since the singular values are the square roots of the eigenvalues of K^*K and by using that $(K^*Kx, x) = \|Kx\|^2$ we get

$$\mu_1 = \max_{\|x\|=1}\|Kx\| \tag{3.10a}$$

$$\mu_{n+1} = \min_{z_1,\ldots,z_n\in\mathcal{X}} \max_{\substack{x\perp z_1,\ldots,z_n\\ \|x\|=1}} \|Kx\|. \tag{3.10b}$$

Note that the right-hand side of (3.10a) is precisely the definition of the operator norm of K, so $\mu_1 = \|K\|$.

By using these relations we can prove the following.

Theorem 3.6 *The singular values of compact operators K_1 and K_2 satisfy*

$$\mu_{n+m+1}(K_1 + K_2) \le \mu_{n+1}(K_1) + \mu_{m+1}(K_2)$$

Proof:

$$\begin{aligned}
\mu_{n+m+1}(K_1 + K_2) &= \min_{z_1,\ldots,z_{n+m}\in\mathcal{X}} \max_{\substack{x\perp z_1,\ldots,z_{n+m}\\ \|x\|=1}} \|(K_1 + K_2)x\| \\
&\le \min_{z_1,\ldots,z_{n+m}\in\mathcal{X}} \max_{\substack{x\perp z_1,\ldots,z_n\\ \|x\|=1}} \|K_1x\| \\
&\quad + \min_{z_1,\ldots,z_{n+m}\in\mathcal{X}} \max_{\substack{x\perp z_{n+1},\ldots,z_{n+m}\\ \|x\|=1}} \|K_2x\| \\
&\le \min_{z_1,\ldots,z_n\in\mathcal{X}} \max_{\substack{x\perp z_1,\ldots,z_n\\ \|x\|=1}} \|K_1x\| \\
&\quad + \min_{z_{n+1},\ldots,z_{m+n}\in\mathcal{X}} \max_{\substack{x\perp z_{n+1},\ldots,z_{n+m}\\ \|x\|=1}} \|K_2x\| \\
&\le \mu_{n+1}(K_1) + \mu_{m+1}(K_2)
\end{aligned}$$

□

If the inverse operator K^{-1} exists and is compact, we must have that K has a finite dimension N (otherwise the inverse is unbounded) and we must have that the null-space is trivial, i.e. that all singular values are non-zero. In this case, it is straightforward to show that the inverse has the following appearance

$$K^{-1}g = \sum_{n=1}^{N}\frac{1}{\mu_n}(g, u_n)v_n.$$

We can now introduce the *condition number* of an operator as

$$\kappa(K) = \|K\|\|K^{-1}\| = \mu_1\frac{1}{\mu_N} = \frac{\mu_1}{\mu_N}, \tag{3.11}$$

since μ_N^{-1} must be the largest singular value of K^{-1}. The condition number is often used in numerical linear algebra to illustrate the "ill-posedness" of certain problems. More on this in Chapter 5, "Ill-posed Problems and Regularization".

What is the adjoint operator in terms of the singular values and -vectors? Let the singular value expansion of $K : \mathcal{X} \mapsto \mathcal{Y}$ be

$$Kf = \sum_{n=1}^{\infty} \mu_n (f, v_n) u_n$$

and let

$$f = \sum_{n=1}^{\infty} (f, v_n) u_n + Qf,$$

where Q projects onto the null-space of K and let

$$g = \sum_{m=1}^{\infty} (g, u_m) u_m + Pg,$$

where P projects onto $\{u_1, u_2, \ldots\}^{\perp}$. We now get

$$\begin{aligned}
(Kf, g) &= \left(\sum_{n=1}^{\infty} \mu_n (f, v_n) u_n, \sum_{m=1}^{\infty} (g, u_m) u_m + Pg \right) \\
&= \sum_{n=1}^{\infty} \mu_n (f, v_n) \left[\overline{(g, u_n)} + (u_n, Pg) \right] \\
&= \sum_{n=1}^{\infty} \mu_n \overline{(g, u_n)} \left[(f, v_n) + (Qf, v_n) \right] \\
&= \left(\sum_{n=1}^{\infty} (f, v_n) v_n + Qf, \sum_{m=1}^{\infty} \mu_m (g, u_m) v_m \right) = (f, K^* g)
\end{aligned}$$

We have here used that $(u_n, Pg) = 0$ from the definition of P and that $(Qf, v_n) = \left(Qf, \frac{1}{\mu_n} K^* u_n \right) = \frac{1}{\mu_n} (KQf, u_n) = 0$. From this we see that

$$K^* g = \sum_{n=1}^{\infty} \mu_n (g, u_n) v_n.$$

3.3 Relating Eigenvalues and Singular Values

Given a normal operator K, there is a straightforward relation between the singular values and the eigenvalues. This is seen by rewriting the spectral decomposition for a normal operator

$$Kf = \sum_{n=1}^{\infty} \lambda_n (f, \varphi_n) \varphi_n = \sum_{n=1}^{\infty} |\lambda_n| (f, \varphi_n) \left(\frac{\lambda_n}{|\lambda_n|} \varphi_n \right),$$

and noticing that this last expression is on singular value expansion form. So for normal (and self-adjoint) operators, we have $\mu_n = |\lambda_n|$ for all n.

The next theorem provides us with some inequalities relating singular values and eigenvalues.

Theorem 3.7 *Let K be a compact linear operator, let (λ_n) denote the eigenvalues and (μ_n) the singular values, ordered such that*

$$|\lambda_1| \geq |\lambda_2| \geq \cdots \quad \textit{and} \quad \mu_1 \geq \mu_2 \geq \cdots .$$

Then

$$\prod_{n=1}^{m} |\lambda_n| \leq \prod_{n=1}^{m} \mu_n \tag{3.12}$$

and

$$\sum_{n=1}^{m} |\lambda_n|^{\alpha} \leq \sum_{n=1}^{m} \mu_n^{\alpha}, \qquad \sum_{n=1}^{\infty} |\lambda_n|^{\alpha} \leq \sum_{n=1}^{\infty} \mu_n^{\alpha} \tag{3.13}$$

for any constant $\alpha > 0$ and all $m \in \mathbb{N}$.

This theorem was proved by Hermann Weyl [Wey49] for finite dimensional mappings, but states that the proof can be extended to "completely continuous linear operators", another word for compact linear operators.

Theorem 3.8 *If there are an infinite number of non-zero eigenvalues we have*

$$\mu_n = \mathcal{O}(n^{-p}) \quad \Rightarrow \quad |\lambda_n| = \mathcal{O}(n^{-p}) \tag{3.14}$$

for positive p.

Proof: Assume $\mu_n = \mathcal{O}(n^{-p})$, that is, there exists a $c > 0$ such that $\mu_n \leq cn^{-p}$ for all n. We now get from the relation (3.12),

$$\prod_{n=1}^{m} |\lambda_n| \leq \prod_{n=1}^{m} \mu_n \leq \prod_{n=1}^{m} cn^{-p} \leq c^m (m!)^{-p}.$$

When inserting the Stirling approximation of the factorial function[1]

$$m! = \sqrt{2\pi m} e^{-m} m^m \Theta(m), \qquad \Theta(m) = 1 + \mathcal{O}\left(\frac{1}{m}\right),$$

we get the following inequality

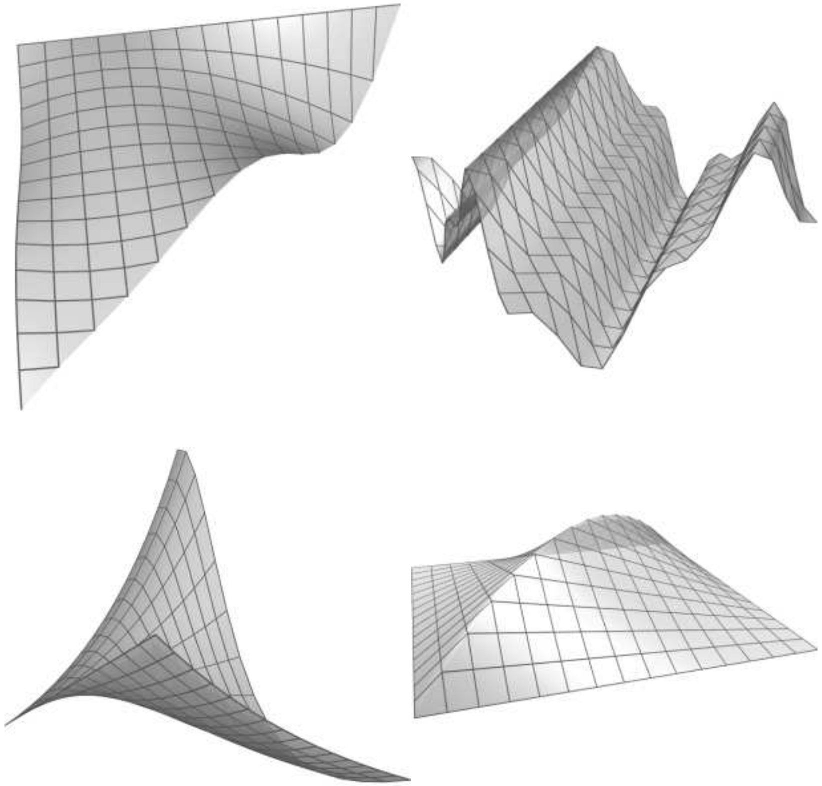
$$\prod_{n=1}^{m} |\lambda_n| \leq (2\pi m)^{-\frac{p}{2}} \left(ce^{-p}\right)^m \left(m^{-p}\right)^m \Theta(m)^{-p}.$$

By setting $c' = ce^{-p}$ we get

$$\prod_{n=1}^{m} \frac{|\lambda_n|}{c' m^{-p}} \leq (2\pi m)^{-\frac{p}{2}} \Theta(m)^{-p} \quad \rightarrow \quad 0,$$

as $m \to \infty$. From this we see that we must have $|\lambda_m| \leq c' m^{-p}$ for $m \to \infty$. $\square$

[1] See [Knu97] page 115 for more information on this approximation.



CHAPTER 4

Operator Classes

class, *to arrange, group, or rate according to qualities or characteristics*
— THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE

The majority of this chapter will address the question: Which operators have eigenvalues and if so, how do they asymptotically behave? An attempt will be made to classify operators that have similar properties with respect to this question.

4.1 Integral Operators

A large part of compact operators can be expressed as integral operators and special focus will be put these.

An integral operator K is generically of the form

$$Kf(\mathbf{s}) = \int_J k(\mathbf{s},\mathbf{t})f(\mathbf{t})d\mathbf{t}, \quad \mathbf{s} \in I. \tag{4.1}$$

The functional $k : I \times J \mapsto \mathbb{C}$ is called the kernel and we say that K is the integral operator induced by k. In the following, unless stated otherwise, we assume that I and J are intervals of the form $[a,b]$, $[a,\infty[$, $\mathbb{R}$ or product spaces of such.

In order to look at operators mapping $L^2(J)$ into $L^2(I)$, the kernel can be quite unrestricted. However, the domain of K, $\mathcal{D}(K)$, has to specified in order to make the integral in (4.1) make sense. This is done by the following definition

$$\mathcal{D}(K) = \Big\{ f \in L^2(J)\ |k(\mathbf{s},\cdot)f \in L^1(J) \text{ for almost every } \mathbf{s} \in I \\ \text{and } g(\mathbf{s}) = \int_J k(\mathbf{s},\mathbf{t})f(\mathbf{t})d\mathbf{t} \ \Rightarrow\ g \in L^2(I) \Big\}. \tag{4.2}$$

We define the range of K as $\mathcal{R}(K) = K(\mathcal{D}(K))$.

This class of integral operators is big and we will, for one, only look at those operators which are closed and bounded. This is ensured by requiring that $\mathcal{D}(K) = L^2(J)$ (Theorem 3.10 in [HS78]).

The key property of the operators in this thesis is their compactness, and a large class of compact operators are integral operators. But there are integral operators that are not compact and there are compact operators that are not integral. See Appendix B for examples of both kinds.

4.2 Hilbert–Schmidt Operators

Hilbert–Schmidt operators represent a class of operators that turn out to be both integral and compact.

Definition 4.1 *An operator $T \in B(\mathcal{V},\mathcal{W})$ is said to be a* ***Hilbert–Schmidt operator*** *if*

$$\sum_{n=1}^{\infty} \|Te_n\|^2 < \infty \tag{4.3}$$

for some orthonormal basis $(e_n) \subset \mathcal{V}$.

Given two orthonormal bases in $\mathcal{V}$, (e_n) and $(\tilde{e}_n)$ we get

$$\begin{aligned}\sum_{n=1}^{\infty} \|Te_n\|^2 &= \sum_{n=1}^{\infty}\sum_{m=1}^{\infty} |(Te_n,\tilde{e}_m)|^2 = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} |(e_n,T^*\tilde{e}_m)|^2 \\ &= \sum_{m=1}^{\infty} \|T^*\tilde{e}_m\|^2\end{aligned}$$

which shows that

$$\sum_{n=1}^{\infty} \|Te_n\|^2 = \sum_{n=1}^{\infty} \|T^*e_n\|^2 = \sum_{n=1}^{\infty} \|T\tilde{e}_n\|^2.$$

This means that when the inequality (4.3) is fulfilled, it is so for every possible basis and we define

$$\|T\|_2 = \left(\sum_{n=1}^{\infty} \|Te_n\|^2\right)^{\frac{1}{2}} \tag{4.4}$$

for an arbitrary basis (e_n). This 2-norm is also called the *Hilbert–Schmidt norm*.

We now consider the case were $T \in B(L^2(J), L^2(I))$. Let (e_n^J) and (e_m^I) be an orthonormal basis for $L^2(J)$ and $L^2(I)$ respectively. The sequence $(e_m^I \otimes \overline{e_n^J})$ will now be an orthonormal basis for $L^2(I \times J)$, see Section A.5 in the appendix. Using Parsevals equality we see from

$$\|Te_n^J\|^2 = \sum_{m=1}^{\infty} |(Te_n^J, e_m^I)|^2 < \infty$$

that the integral kernel k defined as

$$k = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} (Te_n^J, e_m^I) e_m^I \otimes \overline{e_n^J} \tag{4.5}$$

lies in $L^2(I \times J)$. Let K be the integral operator induced by k. We wish to show that this is a bounded operator. By using the Cauchy-

Schwartz inequality we get[1]

$$|Kf(s)| = \left| \int_J k(s,t)f(t)dt \right| \leq \|k(s,\cdot)\| \|f\|,$$

leading to the bound

$$\begin{aligned}\|Kf\|^2 &= \int_I |Kf(s)|^2 ds \leq \int_I \|k(s,\cdot)\|^2 \|f\|^2 ds \\ &= \int_I \int_J |k(s,t)|^2 dt ds \|f\|^2 = \|k\|^2 \|f\|^2,\end{aligned}$$

which shows that K is bounded and $\|K\| \leq \|k\|$. Notice that the kernel norm and operator norm are not, in general, equal.

Now we get

$$\begin{aligned}(Ke_n^J, e_m^I) &= \left(\int_J \sum_{m'=1}^{\infty} \sum_{n'=1}^{\infty} (Te_{n'}^J, e_{m'}^I) e_{m'}^I(\cdot) \overline{e_{n'}^J(t)} e_n^J(t) dt, e_m^I \right) \\ &= \sum_{m'=1}^{\infty} \sum_{n'=1}^{\infty} (Te_{n'}^J, e_{m'}^I) \left(e_{m'}^I \otimes \overline{e_{n'}^J}, e_m^I \otimes \overline{e_n^J} \right) = (Te_n^J, e_m^I).\end{aligned} \tag{4.6}$$

This obviously implies that $Ke_n^J = Te_n^J$ for all n, which shows for arbitrary $x \in L^2(J)$

$$Kx = K \left(\sum_{n=1}^{\infty} (x, e_n^J) e_n^J \right) = \sum_{n=1}^{\infty} (x, e_n^J) K e_n^J = \sum_{n=1}^{\infty} (x, e_n^J) T e_n^J = Tx,$$

because of the continuity of both T and K (they are linear and bounded), so $T = K$. This means that every Hilbert–Schmidt operator is an integral operator.

[1]From the *Fubini Theorem* (see [Rud66]) we have that $k \in L^2(I \times J)$ implies that the function $t \to |k(s,t)|^2$ lies in $L^1(J)$. This means that $\int_J |k(s,t)|^2 dt < \infty \Leftrightarrow k(s,\cdot) \in L^2(J)$.

We now define an integral operator $\widetilde{K}_{MN}$ with kernel

$$\widetilde{k}_{MN} = \sum_{m=1}^{M} \sum_{n=1}^{N} (Te_n^J, e_m^I) e_m^I \otimes \overline{e_n^J}.$$

The operator $\widetilde{K}_{MN}$ has finite rank for every M and N, and clearly $\widetilde{k}_{MN} \to k$ as $M, N \to \infty$. Because $k - \widetilde{k}_{MN}$ induces a well-defined (Hilbert–Schmidt) integral operator we have $\|K - \widetilde{K}_{MN}\| \leq \|k - \widetilde{k}_{MN}\|$ and so $\widetilde{K}_{MN} \to K$ in the operator norm. This shows that K, and hence T, is compact (see Theorem 2.15).

In connection with defining the 2-norm above, see Equation (4.4), it was shown that the norm was invariant with respect to the basis. If we now use the (v_n)-basis from the singular value expansion we get

$$\|K\|_2 = \left(\sum_{n=1}^{\infty} \|Kv_n\|^2 \right)^{\frac{1}{2}} = \left(\sum_{n=1}^{\infty} \|\mu_n u_n\|^2 \right)^{\frac{1}{2}} = \left(\sum_{n=1}^{\infty} \mu_n^2 \right)^{\frac{1}{2}}.$$

We shall later see how this way of defining a norm can be generalized.

What does the adjoint of a (Hilbert–Schmidt) integral operator look like? Let $k \in L^2(I \times J)$ be the kernel of K, $f \in L^2(J)$ and $g \in L^2(I)$. We now get

$$\begin{aligned}(Kf, g) &= \int_I (Kf)(s) \overline{g(s)} ds \\ &= \int_I \left(\int_J k(s,t) f(t) dt \right) \overline{g(s)} ds \\ &= \int_J f(y) \left(\int_I k(s,t) \overline{g(s)} ds \right) dt \\ &= \int_J f(y) \overline{\left(\int_I \overline{k(s,t)} g(s) ds \right)} dt = (f, K^* g)\end{aligned} \tag{4.7}$$

So the kernel of the adjoint operator K^* is the conjugate transposed kernel, $\overline{k(t,s)}$. Such a kernel is called Hermitian.

Now if two integral operators K_1 and K_2 are given, what is the kernel of K_1K_2? Let $k_1 \in L^2(I_0 \times I_1)$ be the kernel for K_1 and $k_2 \in L^2(I_1 \times I_2)$ the kernel for K_2. Then we get

$$\begin{aligned}(K_1K_2f)(s) &= \int_{I_1} k_1(s,z)\int_{I_2} k_2(z,t)f(t)dtdz \\ &= \int_{I_2}\left[\int_{I_1} k_1(s,z)k_2(z,t)dz\right] f(t)dt \\ &= \int_{I_2} k_{1,2}(s,t)f(t)dt.\end{aligned} \tag{4.8}$$

So K_1K_2 is just another integral operator with kernel $k_{1,2} \in L^2(I_0 \times I_2)$.

Let us now consider K^n with associated kernel $k_n \in L^2(I \times I)$. Using the above calculations recursively we get

$$k_1 = k, \qquad k_n(s,t) = \int_I k(s,z)k_{n-1}(z,t)dz. \tag{4.9}$$

The kernels k_n are called the *iterated kernels*.

We shall also define the *trace* of an integral operator K with kernel $k \in L^2(I \times I)$ as:

$$\mathrm{tr}(K) = \int_I k(t,t)dt,$$

whenever it makes sense. The traces of K^n will be called the *higher order traces*.

4.3 Operators Without Surprises

The word "surprise" refers to the eigenvalues. As already seen, self-adjoint operators behave very nicely when it comes to spectral decomposition. The eigenvalues are real and the eigenvectors are mutually orthogonal and form a basis of the range of the operator. But other operators are similarly nice. This section will present some.

4.3.1 Normal Operators

Normal operators possess a full spectral decomposition like self-adjoint operators do. The only difference lies in the fact that the eigenvalues can be complex. Furthermore, as shown in Section 3.3, the absolute value of the eigenvalues are equal to the singular values.

When K is an integral operator with kernel $k \in L^2(I \times I)$ we get by using Equation (4.8):

$$\int_I \overline{k(z,s)}k(z,t)dz = \int_I k(s,z)\overline{k(t,z)}dz, \quad \text{for almost all } (s,t) \in I^2,$$

as the condition for an integral operator to be normal. We shall later see a class of integral operators that are normal.

4.3.2 Rotated Self-Adjoint Operators

Consider an operator that fulfills the relation

$$K^* = \alpha K,$$

where $\alpha \in \mathbb{C}$ and $|\alpha| = 1$. We call such an operator a *rotated Hermitian operator*.[2] Because of

$$K = (K^*)^* = (\alpha K)^* = \overline{\alpha}K^* = |\alpha|^2 K,$$

we see the need for demanding $|\alpha| = 1$. When $\alpha = 1$ the operator is self-adjoint by definition and when $\alpha = -1$ the operator is called anti-Hermitian. By observing

$$K^*K = \alpha K^2 \quad \text{and} \quad KK^* = K(\alpha K) = \alpha K^2,$$

we see that K is normal. Let now $H = \sqrt{\alpha}K$ and we have from

$$H^* = (\sqrt{\alpha}K)^* = \overline{\sqrt{\alpha}}K^* = \overline{\sqrt{\alpha}}\alpha K = |\alpha|\sqrt{\alpha}K = H,$$

[2]This term is an invention of the author since the generalization from anti-Hermitian operators with $\alpha = -1$ is straightforward.

that H is self-adjoint. Assume that λ is an eigenvalue of H so $H\varphi = \lambda\varphi$. This implies

$$\sqrt{\alpha}K\varphi = \lambda\varphi \quad \Leftrightarrow \quad K\varphi = \frac{\lambda}{\sqrt{\alpha}}\varphi.$$

So all spectral properties that hold for self-adjoint operators are inherited by rotated Hermitian operators, except that the eigenvalues all lie on the line $\frac{1}{\sqrt{\alpha}}t, t \in \mathbb{R}$, in the complex plane instead of on the real axis.

4.3.3 Degenerate Kernels

A degenerate kernel is of the generic form

$$k(s,t) = \sum_{n=1}^{N} a_n(s)\overline{b_n(t)}, \quad (s,t) \in I \times J. \tag{4.10}$$

The sequences $(a_n)_{n=1}^N$ and $(b_n)_{n=1}^N$ are each assumed to be linearly independent. If they are not, the expression (4.10) can easily be rewritten so that they are linearly independent (resulting in a smaller value of N).

Let now an orthonormal basis for $\operatorname{span}\{a_1, a_2, \ldots, a_N, b_1, b_2, \ldots, b_N\}$ be $(e_n)_{n=1}^{N'}$ where $N \le N' \le 2N$. There then exists matrices $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{N' \times N}$ such that

$$a_n = \sum_{k=1}^{N'} \mathbf{A}_{k,n} e_k \quad \text{and} \quad b_n = \sum_{k=1}^{N'} \mathbf{B}_{k,n} e_k. \tag{4.11}$$

Let K be the integral operator induced by k and consider the eigenvalue problem $K\varphi = \lambda\varphi$:

$$K\varphi = K\left(\sum_{j=1}^{N'}(\varphi, e_j)e_j\right) = \sum_{j=1}^{N'}(\varphi, e_j)Ke_j = \lambda\varphi \quad \Leftrightarrow$$
$$\sum_{j=1}^{N'}(\varphi, e_j)(Ke_j, e_i) = \lambda(\varphi, e_i) \quad \text{for all } i = 1, 2, \ldots, N'.$$

This shows that the original eigenvalue problem is equivalent to the following matrix eigenvalue problem:

$$\mathbf{W}\mathbf{x} = \lambda\mathbf{x},$$

where $\mathbf{W}_{i,j} = (Ke_j, e_i)$ and $\mathbf{x}_i = (\varphi, e_i)$, $\mathbf{W} \in \mathbb{C}^{N' \times N'}$ and $\mathbf{x} \in \mathbb{C}^{N'}$.

Using the expressions in (4.11) for a_n and b_n we get

$$(Ke_j, e_i) = \left(\int_J \sum_{n=1}^{N} a_n(\cdot)\overline{b_n(t)}e_j(t)dt, e_i\right) = \sum_{n=1}^{N}((e_j, b_n)a_n, e_i)$$
$$= \sum_{n=1}^{N} \overline{\mathbf{B}_{j,n}}\mathbf{A}_{i,n} \quad \Rightarrow \quad \mathbf{W} = \mathbf{A}\mathbf{B}^H.$$

So the eigenvalues of the operator K and the matrix $\mathbf{W}$ are identical. Because of the dimensions of $\mathbf{A}$ and $\mathbf{B}$ the matrix $\mathbf{W}$, and K, can have at most N eigenvalues different from zero. Note also that the infinite number of vectors orthogonal to $e_1, e_2, \ldots, e_{N'}$ are all eigenvectors associated with the eigenvalue $\lambda = 0$, i.e. lies in the null-space of K.

See Section D.6 in the appendix for an example of a degenerate operator.

4.4 The Existence of Eigenvalues of Integral Operators

Does an integral operator always have eigenvalues? The answer is no. As we have just seen in the previous sections, some operators are "well-behaved" and we can foresee the existence of eigenvalues. But when an operator is infinite dimensional and non-normal, it becomes less obvious.

This section will provide some classes of operators that have zero, at least one, or a finite number of eigenvalues. Note however, that the presented results are not exhaustive, that is, do not cover every possible integral operator.

4.4.1 Volterra Operators

A Volterra operator is characterized by the fact that its kernel $k \in L^2(I \times I)$ fulfills $k(s,t) = 0$ for all $t > s$. This means that, if $I = [a,b]$, it is representable in the form

$$Kf(s) = \int_a^s k(s,t)f(t)dt, \quad a \le s \le b.$$

Regarding the existence of eigenvalues, we have the following important theorem.

Theorem 4.2 *The only possible eigenvalue for a Volterra operator is* 0.

Proof: The proof will be limited to the case where $k \in L^2([a,b]^2)$ with $-\infty < a < b < \infty$. Since $C([a,b]^2)$ is dense in $L^2([a,b]^2)$ it is sufficient to look at the case

$$\int_a^s k(s,t)\varphi(t)dt = \lambda\varphi(s), \quad s \in [a,b] \tag{4.12}$$

with $k \in C([a,b]^2)$ and $\varphi \in L^2([a,b])$. Since k is continuous and its domain is a closed and bounded set, there exists an $M > 0$ such that $|k(s,t)| \le M$ for all $(s,t) \in [a,b]^2$.

Assume that $\lambda \neq 0$ is an eigenvalue with corresponding eigenvector φ. We now have

$$\|\varphi\|_1 = \int_a^b |\varphi(t)| \cdot 1dt \le \|\varphi\|_2\|1\|_2 = \sqrt{b-a}\|\varphi\|_2 < \infty \tag{4.13}$$

according to Cauchy-Schwartz' inequality, since $1_{[a,b]} \in L^2([a,b])$.

We now wish to show that

$$|\varphi(s)| \le |\mu|^n M^n \|\varphi\|_1 \frac{(s-a)^{n-1}}{(n-1)!} \quad \text{for all } s \in [a,b], \tag{4.14}$$

where $\mu = \frac{1}{\lambda}$, holds for all $n \in \mathbb{N}$. This can be done by induction. Since

$$|\varphi(s)| \le |\mu| \int_a^s |k(s,t)||\varphi(t)|dt \le |\mu|M\|\varphi\|_1$$

the inequality (4.14) is shown for $n = 1$. Assume now that (4.14) holds for $n = k-1$. We now get

$$\begin{aligned}
|\varphi(s)| &\le |\mu| \int_a^s |k(s,t)||\varphi(t)|dt \\
&\le |\mu| \int_a^s |k(s,t)||\mu|^{k-1}M^{k-1}\|\varphi\|_1 \frac{(t-a)^{k-2}}{(k-2)!}dt \\
&\le |\mu|^k M^k \|\varphi\|_1 \int_a^s \frac{(t-a)^{k-2}}{(k-2)!}dt = |\mu|^k M^k \|\varphi\|_1 \frac{(s-a)^{k-1}}{(k-1)!}.
\end{aligned}$$

Since (4.14) now holds for all $n \in \mathbb{N}$, we can make the right-hand side arbitrarily small. This implies that $\varphi(s) = 0$ for all $s \in [a,b]$.

So no non-trivial solutions can exist for non-zero eigenvalues. □

4.4.2 Positive and Symmetrizable Kernels

Some theorems now follow that tell when an integral operator will have *at least one* eigenvalue.

In [Hoc73], we have the following theorem.

Theorem 4.3 *Let K be an integral operator with kernel $k \in C(I^2)$ where I is closed and bounded. Then K has at least one eigenvalue if and only if*

$$\operatorname{tr}(K^n) \neq 0.$$

for some $n \ge 2$.

A kernel $k \in L^2(I \times J)$ is denoted *positive* if, not surprisingly, $k(s,t) > 0$ for all $(s,t) \in I \times J$. Note that this concept should not be confused with positive operators.[3] The above theorem can then be used to show the following theorem.

Theorem 4.4 *Let K be an integral operator with kernel $k \in C(I^2)$ where I is closed and bounded. If k is positive then K has at least one eigenvalue.*

[3] An operator T is positive if $(Tx, x) > 0$ for all $x \neq 0$.

Proof: Consider the operator K^2, whose kernel is given by Equation (4.9),

$$k_2(s,t) = \int_I k(s,z)k(z,t)dz.$$

This kernel is positive which means that the trace, $\operatorname{tr}(K^2)$, is also positive. By Theorem 4.3 at least one eigenvalues must now exist. $\square$

Results related to positive kernels can also be found in [Coc72]. In this book it is further proved that *symmetrizable* operators have at least one eigenvalue. Such an operator is defined in the following way.

Definition 4.5 *An operator K is (left)* ***symmetrizable*** *if there exists a semidefinite self-adjoint operator H such that HK is both self-adjoint and non-null.*

The usefulness of this class is unclear, since no example could be found of a symmetrizable operator that was non-normal and non-degenerate.

4.4.3 On Operators with only a Finite Number of Eigenvalues

Identifying integral operators with only a finite number of eigenvalues can be very relevant. This has been the subject of some papers. In [Swa71], the following curious theorem can be found.

Theorem 4.6 *An operator K with kernel $k \in L^2$ possesses precisely q non-zero eigenvalues if and only if the higher order traces $k^{(n)} = \operatorname{tr}(K^n)$, $n \geq 2$, satisfy a recurrence relation of the form*

$$k^{(q+n)} + \mathbf{a}_1 k^{(q+n-1)} + \cdots + \mathbf{a}_{q-1} k^{(n+1)} + \mathbf{a}_q k^{(n)} = 0, \quad \textit{for all } n \geq 2$$

where $\mathbf{a} \in \mathbb{C}^q$ is vector with $\mathbf{a}_q \neq 0$.

The result, in the words of the author, seems to be *"of more theoretical than practical interest."* But it seems to suggest that characterizing operators that only possess a finite number of eigenvalues is not a simple subject.

The above theorem deals only with the (higher order) traces of an integral operator. What if the trace is not well-defined because of some discontinuity along the diagonal? In [Swa71] the author states that an arbitrary L^2 kernel can be modified and redefined on the diagonal without changing the essence of the integral equation, in particular without changing the eigenvalues.

An article by Samuel Karlin [Kar64] also focuses on the existence of eigenvalues for integral operators. One of the results is very interesting. He introduces the kernel class *extended totally positive*, abbreviated ETP. A kernel $k \in C^\infty([a,b]^2)$ is said to be ETP if

$$\det\left(\left[\frac{\partial^{i+j}k(s,t)}{\partial s^i \partial t^j}\right]_{i,j=0}^n\right) > 0$$

for all $n = 0, 1, 2, \ldots$ and all $(s,t) \in [a,b]^2$. He now proves

Theorem 4.7 *Let a kernel $k \in C^\infty([a,b]^2)$ be of class ETP. Then the induced integral operator K possesses a countable set of simple positive eigenvalues*

$$\lambda_1 > \lambda_2 > \lambda_3 > \cdots$$

decreasing to zero. Let φ_n denote the corresponding eigenfunctions. Then

$$\det\left([\varphi_i(x_j)]_{i=0,j=1}^{n-1,n}\right) > 0 \quad (< 0) \tag{4.15}$$

for all choices of $n \in \mathbb{N}$ and $a \leq x_1 < x_2 < \cdots < x_n \leq b$.

The Equation (4.15) may seem a bit strange, but Karlin states that it implies that φ_n has precisely n zeros and zeros of successive eigenfunctions strictly interlace. This is a very strong characterization of the eigenfunctions that shows how they become more and more oscillatory.[4]

The ETP kernel class may seem somewhat abstract. A concrete class of kernels that lie in ETP are mentioned to be

$$k(s,t) = \sum_{n=0}^{\infty} a_n \left[x(s)\right]^n \left[y(t)\right]^n$$

where $a_n \geq 0$ for all $n \in \mathbb{N}$ and strictly positive for infinitely many n.

4.5 Singular Values and Eigenvalues of Integral Operators

The journal *Acta Mathematica* published in 1931 an article by E. Hille and J. D. Tamarkin entitled *"On the Characteristic Values of Linear Integral Equations"* (see [HT31]). The main question was, in the words of themselves: *"What can be said about the distribution of the characteristic values of the Fredholm integral equation [...] on the basis of the general analytic properties of the kernel [...] such as integrability, continuity, differentiability, analyticity and the like?"*

The article included a table in which their results were summarized. This table is re-created in Appendix C on page 141.

Their work lay a solid foundation that is still referenced to today. Some of the results were later improved by various people, and especially bounds for singular values, which the Hille and Tamarkin article did not cover, have emerged.

The following sections will summarize some of the bounds that can be found in the literature.

[4]This is very important when talking regularization, since regularized solution often are assembled from eigenvectors (singular vectors) corresponding to the largest eigenvalues (singular values).

4.5.1 Analytic Kernels

A function is said to be analytic if it is locally representable as a convergent power series. It is a strong property that is stronger than being infinitely many times continuously differentiable (see [MH87, page 205] for more information).

The eigenvalues of operators with analytic kernels will decay exponentially. Hille and Tamarkin showed that the asymptotic decay was

$$|\lambda_n| = \mathcal{O}(R^{-\frac{1}{4}n})$$

where the constant R is related to how big an area the kernel is analytic in (see the exact result in Appendix C). For Hermitian kernels, this result was improved by Little and Reade [LR84] to $\mathcal{O}(R^{-n})$, where the constant R was defined as in the case above.

The behavior is much the same for singular values. The following theorem is from [Kre99], that actually builds upon the result by Little and Reade.

Theorem 4.8 *Let $K : L^2(I) \mapsto L^2(I)$, where I is a finite interval, be an integral operator with analytic kernel on $I \times I$. Then the singular values of K decay at least exponentially $\mu_n(A) = \mathcal{O}(R^{-n})$ for some constant $R > 1$. Furthermore, the estimate cannot be improved to $\mathcal{O}(R^{-(1+\epsilon)n})$ for any $\epsilon > 0$.*

Proof: We will only outline the proof, see [Kre99] for details. First, by proper scaling, it is sufficient to consider the interval $I = [-1, 1]$. Let the kernel of K be denoted k and define the kernel of the operator K_n as

$$k_n(s,t) = \frac{1}{2}T_0(s)a_0(t) + \sum_{m=1}^{n} T_m(s)a_m(t)$$

where $T_m(\cos\theta) = \cos(m\theta)$ is the mth Chebyshev polynomial and $a_m : \mathbb{R} \mapsto \mathbb{C}$ are coefficient functions. We have $K_n \to K$ for $n \to \infty$ and it can be proven that $\|K - K_n\| = \mathcal{O}(R^{-n})$ for some $R > 1$. Now we get from Theorem 3.6 that

$$\mu_{n+2}(K) \leq \mu_1(K - K_n) + \mu_{n+2}(K_n) = \|K - K_n\| = \mathcal{O}(R^{-n}),$$

where $\mu_{n+2}(K_n) = 0$ because the dimension of the range of K_n is at most $n+1$ and therefore K_n has no more than $n+1$ singular values different from zero. □

A subset of analytic functions are the *entire* functions, which are analytic in the entire complex plane. For such kernels, the decay of the eigenvalues is even faster. Some kernels will actually lead to an asymptotic decay of $|\lambda_n| = \mathcal{O}(n^{-\alpha n})$, see Section D.2 in the appendix for an example.

4.5.2 Discontinuous Derivatives

It turns out that it is a discontinuity in some partial derivative of the kernel with respect to s that is important. For simpler expressions, let us introduce the notation $\partial_s^n k = \frac{\partial^n}{\partial s^n} k$ as the nth partial derivative of k with respect to s. Whereas analytic kernels lead to exponential decay of the eigenvalues, the decay is only polynomial when $\partial_s^n k$ is discontinuous for some n.

A special case of the result for kernels in the class Υ_b (see Appendix C for the definition of this class) is the following.[5]

Theorem 4.9 *Let an integral kernel $k \in L^2([a,b]^2)$ and let the partial derivatives $\partial_s^n k(s,t)$ be continuous for $n = 1, 2, \ldots, v-1$, $v \geq 0$. Let furthermore the vth partial derivative fulfill*

$$\partial_s^v k(s,t) = \int_a^s g(z,t)dz + C(t),$$

for some function C and where

$$\int_a^b \left[\int_a^b |g(s,t)|^p ds \right]^{\frac{1}{p-1}} dt < \infty$$

[5] This special case is also mentioned in [Coc72].

for $1 < p \leq 2$. Then the eigenvalues of the induced integral operator behave asymptotically as

$$|\lambda_n| = \mathcal{O}\left(n^{-(v+2-\frac{1}{p})} \right).$$

Let us look at the class that Hille and Tamarkin called $\mathrm{Lip}_3(v)$ where $v \geq 1$. A kernel $k \in L^2([0,2\pi]^2)$ lies in this class if the partial derivatives

$$\partial_s^n k(s,t), \quad n = 1, 2, \ldots, v-1$$

all exist and are continuous in s for fixed t. Furthermore, the function $\partial_s^v k(s,t)$, considered as a periodic function of s outside the interval $[0,2\pi]$, must satisfy the condition

$$\int_0^{2\pi} |\partial_s^v k(s+\epsilon,t) - \partial_s^v k(s,t)| \, ds < g(t),$$

where $g \in L^\infty$, i.e. is bounded, and where $\epsilon > 0$ is sufficiently small. Since

$$\begin{aligned} \int_0^{2\pi} & |\partial_s^v k(s+\epsilon,t) - \partial_s^v k(s,t)| \, ds \\ & \leq \int_0^{2\pi} |\partial_s^v k(s+\epsilon,t)| \, ds + \int_0^{2\pi} |\partial_s^v k(s,t)| \, ds \\ & = 2\int_0^{2\pi} |\partial_s^v k(s,t)| \, ds = 2\|\partial_s^v k(\cdot,t)\|_1, \end{aligned}$$

we see that if $\|\partial_s^v k(\cdot,t)\|_1$ is bounded for all t, then $k \in \mathrm{Lip}_3(v)$. So now we can formulate a simpler theorem concerning kernels that have a discontinuous partial derivative.

Theorem 4.10 *Let an integral kernel $k \in L^2([0,2\pi]^2)$ and let the partial derivatives $\partial_s^n k(s,t)$ be continuous for $n = 1, 2, \ldots, v-1$, $v \geq 0$. If furthermore the vth partial derivative fulfills that $\|\partial_s^v k(\cdot,t)\|_1$ is bounded for all t, then the eigenvalues of the induced integral operator behave asymptotically as*

$$|\lambda_n| = \mathcal{O}\left(n^{-v} (\log n)^{v+\frac{1}{2}} \right).$$

In the article by Hille and Tamarkin, they are not satisfied with the presence of the logarithms in the expressions for the Lip-classes, rows 6, 9 and 10 in the table. They write: *"It is very probable, however, that the presence of the logarithmic factors in the estimates [...] is due to the imperfection of the method used, and that actually these factors should be removed or even replaced by logarithmic factors with exponents of opposite signs."*

Practical use of both Theorem 4.9 and 4.10
can be seen in the appendix, Section D.4.

When it comes to singular values, Smithies [Smi37] proved in 1937 the following theorem.

Theorem 4.11 *Let the integral operator K have a kernel $k \in L^2([0,\pi]^2)$ and let the partial derivatives $\partial_s^n k(s,t)$ be absolutely continuous for $n = 1, 2, \ldots, v-1$, $v \geq 0$. Let furthermore $\partial_s^v k(\cdot,t) \in L^p([0,\pi])$ for almost all t and for $1 < p \leq 2$. If we also have*

$$\int_0^\pi \left[\int_0^\pi |\partial_s^n k(s+\epsilon,t) - \partial_s^n k(s-\epsilon,t)|^p \, ds\right]^{\frac{2}{p}} dt \leq A|\epsilon|^{2\alpha}$$

for some A and for all sufficiently small ϵ, where either $r > 0$, $\alpha > 0$ or $r = 0$, $\alpha > \frac{1}{p} - \frac{1}{2}$, then the eigenvalues of the operator K fulfills

$$\mu_n = \mathcal{O}\left(n^{-v+\alpha+1-\frac{1}{p}}\right).$$

4.5.3 Class $\mathcal{C}_p$ operators.

A slightly different approach to the asymptotic decay of eigenvalues and singular values can be made by introducing special operator norms. These can be defined for an operator K in terms of its singular values as

$$\|K\|_p = \left(\sum_{n=1}^{\infty} \mu_n^p\right)^{\frac{1}{p}}, \tag{4.16}$$

for all $p \geq 1$. Note that for $p = 2$ this norm is equivalent to the Hilbert–Schmidt norm defined in Equation (4.3).

Using this, we see that $\|K\|_\infty = \sup_{n\in\mathbb{N}} |\mu_n| = \mu_1 = \|K\|$ (for the last equality, see Equation (3.10) and the comment that follows).

We now define $\mathcal{C}_p$ as the set of compact operators K for which $\|K\|_p < \infty$. Note that $\mathcal{C}_2$ is equivalent to the Hilbert–Schmidt class introduced earlier. An operator of the class $\mathcal{C}_1$ is also sometimes termed a *trace class* or *nuclear* operator.

It can be shown that $\mathcal{C}_p \subset \mathcal{C}_q$ whenever $p \leq q$. This means that every integral operator with a kernel in L^2 lies in $\mathcal{C}_p$ for some $p \leq 2$.

Because of inequality (3.13), concerning sums of eigenvalues related to sums of singular values, we see that for a compact operator $K \in \mathcal{C}_p$ with eigenvalues ordered as usual, we have

$$\sum_{n=1}^{\infty} |\lambda_n|^p < \sum_{n=1}^{\infty} \mu_n^p < \infty. \tag{4.17}$$

When considering composite operators, the following theorem can prove useful.

Theorem 4.12 *If $K_1 \in \mathcal{C}_p$ and $K_2 \in \mathcal{C}_q$ we have $K_1K_2 \in \mathcal{C}_r$ where $\frac{1}{r} = \frac{1}{p} + \frac{1}{q}$.*

An application of this theorem could be: Given two Hilbert–Schmidt operators, $K_1, K_2 \in \mathcal{C}_2$, we have that $K = K_1K_2 \in \mathcal{C}_1$ which in turn means that the eigenvalues of K will fulfill the inequality in Equation (4.17) with $p = 1$. Note that this result is also included in the Hille and Tamarkin table.

See [DS63] or [Coc72] for proofs and more information.

4.6 Other Operators

Whereas the previous results have covered very broad classes of operators and kernels, more can often be said when focusing on smaller classes. The following sections will discuss two such classes.

4.6.1 Polar Kernels

Consider an integral operator $K : L^2(I) \to L^2(I)$ with a kernel of the type

$$k(s,t) = A(s)g(s,t)B(t),$$

where $k, g \in L^2(I \times I)$, A and B are real and positive functions and the kernel g is Hermitian. If $A(s) = 1$ everywhere the kernel k is called *polar*, see [Han76, page 235] or [Coc72, page 272].

Consider the eigenvalue problem

$$\int_I k(s,t)\varphi(t)dt = \int_I A(s)g(s,t)B(t)\varphi(t)dt = \lambda\varphi(s), \quad s \in I,$$

and set

$$h(s,t) = \sqrt{A(s)B(s)}g(s,t)\sqrt{A(t)B(t)} \quad \text{and} \quad \theta(t) = \sqrt{\frac{B(t)}{A(t)}}\varphi(t). \tag{4.18}$$

Now we get

$$\begin{gathered}\int_I h(s,t)\theta(t)dt = \lambda\theta(s) \quad \Leftrightarrow \\ \int_I \sqrt{A(s)B(s)}g(s,t)\sqrt{A(t)B(t)}\sqrt{\frac{B(t)}{A(t)}}\varphi(t)dt = \lambda\sqrt{\frac{B(s)}{A(s)}}\varphi(s) \quad \Leftrightarrow \\ \int_I A(s)g(s,t)B(t)\varphi(t)dt = \lambda\varphi(s).\end{gathered}$$

So the eigenvalues of K are identical to those of the integral operator H induced by the kernel h. Since h is an obvious Hermitian kernel, the operator H is self-adjoint and the spectral theorem for compact self-adjoint operators applies. This, for one, ensures that there exists a sequence of orthonormal eigenfunctions that spans the range of H. Notice that these eigenfunctions of H, when transformed appropriately (see Equation (4.18)), are also eigenfunctions of K. They are not guaranteed to be orthonormal, though.

4.6.2 Periodic Difference Kernels

Consider the integral operator

$$Kf(s) = \int_{-\pi}^{\pi} k(s-t)f(t)dt, \tag{4.19}$$

where $k \in L^2([-\pi,\pi])$ is 2π-periodic. Using the following orthonormal basis for $L^2([-\pi,\pi])$,

$$e_n(t) = \frac{1}{\sqrt{2\pi}}e^{int},$$

we have

$$k = \sum_{n=-\infty}^{\infty}(k,e_n)e_n.$$

Consider the following mapping using this representation,

$$\begin{aligned}Ke_n(s) &= \int_{-\pi}^{\pi}\left(\sum_{m=-\infty}^{\infty}(k,e_m)e_m(s-t)\right)e_n(t)dt \\ &= \sum_{m=-\infty}^{\infty}(k,e_m)e_m(s)\int_{-\pi}^{\pi}e_{-m}(t)e_n(t)dt \\ &= \sum_{m=-\infty}^{\infty}(k,e_m)e_m(s)(e_n,e_m) = (k,e_n)e_n(s).\end{aligned}$$

Here we have used that $e_m(s-t) = e_m(s)e_{-m}(t)$ and that $\overline{e_n(t)} = e_{-n}(t)$.

So *every* operator with a 2π-periodic difference kernel has e_n, $n \in \mathbb{Z}$, as eigenvectors and (k,e_n) as the corresponding eigenvalues. This means that K can be written as

$$Kx = \sum_{n=-\infty}^{\infty}(k,e_n)(x,e_n)e_n,$$

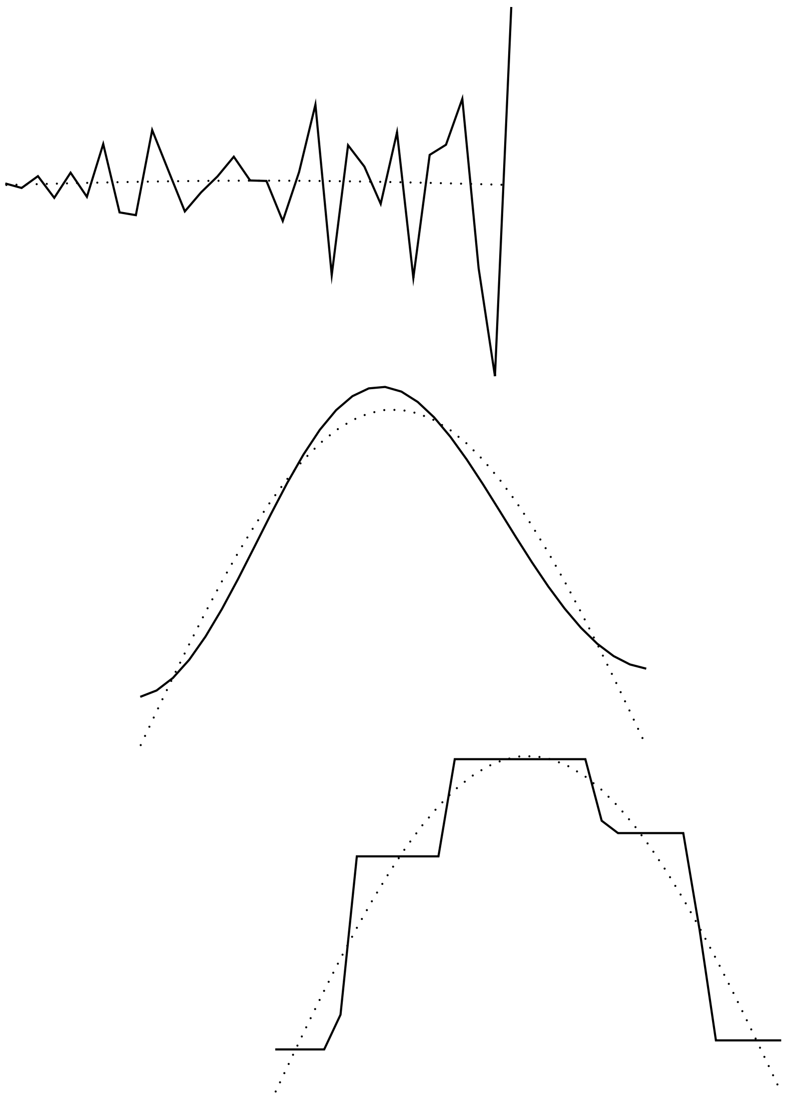
so K is actually normal.

Note that the above conclusions can easily be generalized to operators with kernels $k \in L^2([a,b])$ that are $(b-a)$-periodic.

As a side-remark let us note that similar properties hold for *circulant* matrices. For instance, a $\mathbf{H} \in \mathbb{C}^{4\times 4}$ circulant matrix has the structure

$$\mathbf{H} = \begin{bmatrix} \mathbf{h}_0 & \mathbf{h}_3 & \mathbf{h}_2 & \mathbf{h}_1 \\ \mathbf{h}_1 & \mathbf{h}_0 & \mathbf{h}_3 & \mathbf{h}_2 \\ \mathbf{h}_2 & \mathbf{h}_1 & \mathbf{h}_0 & \mathbf{h}_3 \\ \mathbf{h}_3 & \mathbf{h}_2 & \mathbf{h}_1 & \mathbf{h}_0 \end{bmatrix}$$

for some vector $\mathbf{h} \in \mathbb{C}^4$. The mapping $\mathbf{y} = \mathbf{H}\mathbf{x}$ can now be seen as a discrete periodic convolution. Properties with respect to eigenvalues and other quantities are similar to the continuous case mentioned above, see e.g. [Loa92, Section 4.2].



CHAPTER 5

Ill-Posed Problems and Regularization

ill, *not favorable*
— THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE

regular, *agreeable to an established rule, law, principle, or type*
— WEBSTER'S REVISED UNABRIDGED DICTIONARY

This chapter will present precise definitions of ill-posedness and regularization. This will make it clearer as to why some problems are so "ill".

5.1 Ill-Posed Problems

Ill-posedness can be defined for operators in general:

Definition 5.1 *Let $T : \mathcal{X} \mapsto \mathcal{Y}$ be an operator. Then the equation*

$$Tx = y \tag{5.1}$$

with $y \in \mathcal{Y}$ is called ***well-posed*** *if T is bijective and the inverse operator $T^{-1} : \mathcal{Y} \mapsto \mathcal{X}$ is continuous. Otherwise the equation is called* ***ill-posed****.*

The requirement $y \in \mathcal{Y}$ may seem redundant since the equation is meaningless otherwise. It has been included to stress that well-posedness does not depend only on the operator T but also on the right-hand side y.

So one of the following three factors will make an equation ill-posed:

- T is not surjective. If $y \notin T(\mathcal{X})$, Equation (5.1) may not have a solution (*nonexistence*).
- T is not injective. Equation (5.1) may have more than one solution (*nonuniqueness*).
- T is bijective but T^{-1} is discontinuous. Then the solution does not depend continuously on the right-hand side (*instability*).

Let us consider the case where $\mathcal{X}$ and $\mathcal{Y}$ are *finite dimensional* Hilbert spaces and where $T \in B(\mathcal{X}, \mathcal{Y})$. From Equation (2.12) we know that

$$\mathcal{Y} = \overline{\mathcal{R}(T)} \oplus \mathcal{R}(T)^{\perp} = \overline{\mathcal{R}(T)} \oplus \mathcal{N}(T^*).$$

Since $\mathcal{Y}$ is finite dimensional, the range of T is always closed so $\overline{\mathcal{R}(T)} = \mathcal{R}(T)$. Now given an equation $Tx = y$ we can write the right-hand side as $y = y_{\mathcal{R}} + y_{\mathcal{R}^{\perp}}$ where $y_{\mathcal{R}} \in \mathcal{R}(T)$ and $y_{\mathcal{R}^{\perp}} \in \mathcal{R}(T)^{\perp}$. If $y_{\mathcal{R}^{\perp}} \neq 0$ a solution does not exist (it is ill-posed). Instead, we can seek to find a *least squares solution*

$$\min_{x \in \mathcal{X}} \|y - Tx\|_2. \tag{5.2}$$

By definition there exists an $x \in \mathcal{X}$ such that $Tx = y_{\mathcal{R}}$ and this is obviously a solution to the least squares problem. Since such a solution x fulfills $y - Tx = y_{\mathcal{R}^{\perp}} \in \mathcal{R}(T)^{\perp}$, we have that x solves (5.2) if and only if $y - Tx \in \mathcal{R}(T)^{\perp}$. This is equivalent, since $\mathcal{R}(T)^{\perp} = \mathcal{N}(T^*)$, to

$$T^*(y - Tx) = 0 \quad \Leftrightarrow \quad T^*Tx = T^*y.$$

This new equation is called the *normal equation* associated with $Tx = y$. The solution may not be unique, however, since several solutions may exist that solves $Tx = y_{\mathcal{R}}$. But the solution that has the smallest norm is.[1] This way of choosing a least squares solution to $Tx = y$ is often written $x = T^{\dagger}y$ where $T^{\dagger}$ is called the *Moore–Penrose generalized inverse* of T.

Although the definition of ill-posedness was for operators in general, we will limit ourselves to only look at compact linear operators.

We already know that linear mappings between finite dimensional vector spaces, hence (finite) matrices, are compact. So when are matrix equations well-posed? Obviously square and regular matrices are bijective, and regular matrices always have bounded inverses. Any other case, when the matrix is singular or non-square, will result in a mapping that is not bijective. So only square and regular matrices induce well-posed equations.

But what about infinite dimensional compact operators? This is answered by the following proposition.

Proposition 5.2 *Let $\mathcal{H}$ be an infinite dimensional Hilbert space. If $K \in B(\mathcal{H})$ is compact and K^{-1} exists, then K^{-1} is unbounded.*

Proof: Let (e_n) be an orthonormal sequence in $\mathcal{H}$. Then $Ke_n \to 0$ (see the remarks following Theorem 2.16) but $\|K^{-1}(Ke_n)\| = \|e_n\| = 1$ for all n, so K^{-1} is not continuous. □

Given an infinite dimensional compact operator, *every* equation involving it will be ill-posed. Either the operator is not bijective or the inverse is discontinuous.

[1] The set of solutions to $Tx = y_{\mathcal{R}}$ can be written as $S = \{x_p + x_0 \mid x_0 \in \mathcal{N}(T)\}$ where x_p is a particular solution, $Tx_p = y_{\mathcal{R}}$. Since $\mathcal{N}(T)$ is a subspace it is convex, and translating the convex set $\mathcal{N}(T)$ by x_p does not change the convexity. So S is nonempty, closed and convex and this implies that it contains a *unique* element of smallest norm (Theorem 4.10 in [Rud66]).

5.2 Operator Smoothing and the Picard Condition

One of the implications of Theorem 2.16 was that $Ke_n \to 0$ for compact K and orthonormal (e_n). If K is an integral operator with real kernel $k \in L^2([0,\pi] \times [0,\pi])$ and we consider the usual sine basis, we get

$$\int_0^\pi k(\cdot,t)\sin(nt)dt \to 0, \quad \text{for } n \to \infty,$$

with convergence in the L^2 sense. This is also known as the Riemann-Lebesgue lemma. This clearly shows how high-frequency components are damped by K.

Theorem 5.3 (The Picard Condition) *Let $K \in B(\mathcal{X},\mathcal{Y})$ be a compact linear operator with singular values (μ_n) and corresponding singular vectors $(v_n) \subset \mathcal{X}$ and $(u_n) \subset \mathcal{Y}$. The equation*

$$Kf = g \tag{5.3}$$

is solvable if and only if $g \in \overline{\mathcal{R}(K)}$ and

$$\sum_{n=1}^{\infty} \frac{|(g,v_n)|^2}{\mu_n^2} < \infty. \tag{5.4}$$

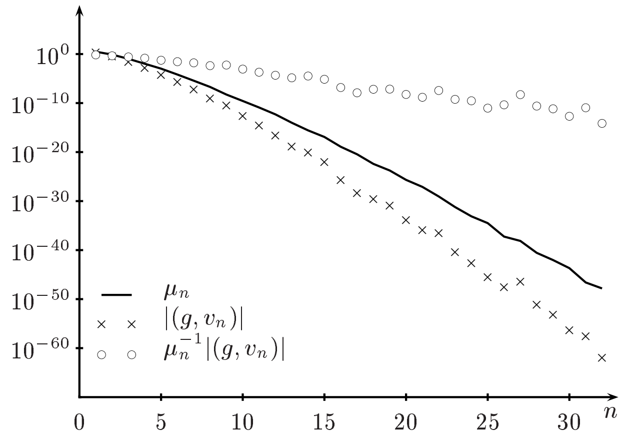
A solution is then given by

$$f = \sum_{n=1}^{\infty} \frac{(g,v_n)}{\mu_n} u_n. \tag{5.5}$$

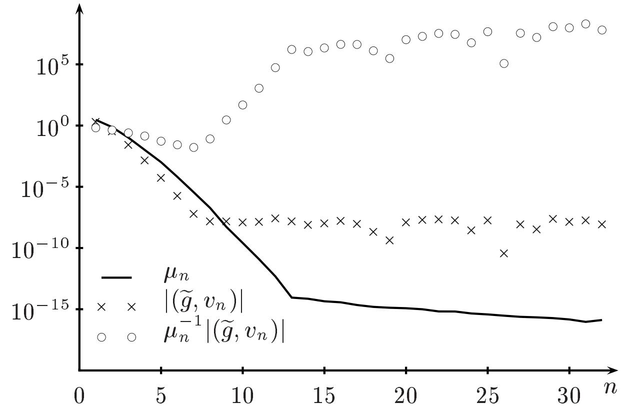
Proof: See [Kre99, page 279]. □

Note that the inequality in (5.4) can easily be fulfilled even if g has components in the null-space of K. This is why the condition $g \in \overline{\mathcal{R}(K)}$ is needed.

Consider now a model problem $Kf = g$ involving a compact operator K. Figure 5.1(a) shows the key quantities in the Picard condition,



(a) A possible Picard plot in infinite precision.



(b) The way the Picard plot above could look on a finite precision machine and with a noisy right-hand side $\widetilde{g}$.

Figure 5.1: *An example of a Picard plot.*

μ_n, $|(g, v_n)|$ and $\mu_n^{-1}|(g, v_n)|$, and we now assume that we wish to compute the solution using expression (5.5).

The singular values μ_n are seen to decrease approximately exponentially. The right-hand side Fourier coefficients, $|(g, v_n)|$, decrease similarly but quicker, so that the quantities $\mu_n^{-1}|(g, v_n)|$ also *decrease* exponentially. This means, assuming they continue to decrease, that the solution will fulfill Equation (5.4) and so a solution will exist and be computable (in theory).

Consider now the plot in Figure 5.1(b). The equation is supposed to be the same as before, but we now have several "noise sources". We now work on a finite precision machine and noise has been added to the right-hand side producing $\widetilde{g} = g + e$ where e is some noise vector. Furthermore, the operator has been approximated by a 32×32 matrix and the right-hand side has similarly been approximated by a 32 element vector.

This introduces some problems. Firstly, the singular values[2] stop to decrease around 10^{-14} because of rounding errors on the machine that computed these values. Secondly, the right-hand side Fourier coefficients decrease as before, but level out around 10^{-8} due to noise in the right-hand side.

This means that the summation formula expressing the solution, see Equation (5.5), can be seperated into two parts. When $n \leq 7$, all quantities are influenced by relatively little noise. When $n > 7$, noise influences either $|(\widetilde{g}, v_n)|$ or μ_n and each term becomes unreliable.

All in all, the solution will be dominated by noise. Note that the noisy model problem was finite dimensional and all singular values were non-zero. This means the the Picard condition is trivially fulfilled. But finite precision and noise in the data makes the solution useless. For information on a so-called discrete Picard Condition and on dealing with noisy right-hand sides, see e.g. [Han98a].

The above example illustrates some of the problems when dealing with ill-posed problems. Even if we do not try to solve the problem

[2]In this discussion, the differences between the singular values of the 32×32 matrix and the 32 largest eigenvalues of the operator K are assumed negligible. The next chapter provides more accurate results on these approximation errors.

using Formula (5.5), solution methods typically have the fact in common, that even if the problem is finite dimensional and the inverse K^{-1} exists, there are two main problems:

- We are not interested in $K^{-1}\widetilde{g}$ but $K^{-1}g$, that is, we want the true solution and not the one influenced by noise.
- Rounding errors tend to blow up somewhere in the solution process.

5.3 Regularization

A typical way to handle these problems is called regularization. Regularization can be defined in a general way as follows.

Definition 5.4 *Let $\mathcal{X}$ and $\mathcal{Y}$ be normed spaces and let $T : \mathcal{X} \mapsto \mathcal{Y}$ be an injective bounded linear operator. Then a family of bounded linear operators $T_\alpha^\# : \mathcal{Y} \mapsto \mathcal{X}$, $\alpha > 0$, with the property of pointwise convergence*

$$\lim_{\alpha \to 0} T_\alpha^\# T x = x, \quad x \in \mathcal{X}, \tag{5.6}$$

is called a ***regularization scheme*** *for the operator T. The parameter α is called the* ***regularization parameter****.*

Note that the limit in Equation (5.6) could just as well be for $\alpha \to \infty$ or with α always integer etc. This would not change the meaning of the definition or the following discussion.

Now assume we wish to solve $Tx = y$ but all we have available is an perturbed right-hand side y^δ such that

$$\|y^\delta - y\| \leq \delta$$

for some error level δ. As mentioned earlier, we are not interested in $T^{-1}y^\delta$ (if the inverse exists), but $T^{-1}y$. The hope is to find an appropriate regularization scheme such that for some $\alpha > 0$ we have $T_\alpha^\# y^\delta \simeq T^{-1}y$.

For a fixed α, let the regularized solution be $x_\alpha^\delta = T_\alpha^\# y^\delta$. The approximation error is then

$$x_\alpha^\delta - x = T_\alpha^\# y^\delta - T_\alpha^\#(y - Tx) - x = T_\alpha^\#(y^\delta - y) + T_\alpha^\# Tx - x,$$

and by using the triangle inequality we get

$$\|x_\alpha^\delta - x\| \le \delta \|T_\alpha^\#\| + \|T_\alpha^\# Tx - x\|.$$

We would now like to make both $\|T_\alpha^\#\|$ small and $T_\alpha^\# T \simeq I$. Unfortunately, when dealing with infinite dimensional compact operators, the following can be shown ([Kre99, p. 269]):

- The operators $T_\alpha^\#$ cannot be uniformly bounded with respect to α, that is, a constant C so that $\|T_\alpha^\#\| \le C$ for all $\alpha > 0$ does *not* exist.
- The operators $T_\alpha^\# T$ do *not* converge in norm to the identity as $\alpha \to 0$.[3]

A popular way of putting this: For ill-posed problems in general the perfect regularization scheme and -parameter does not exist.

One possible regularization scheme is to use the solution expressed by (5.5) but using only the first k terms of the sum. This way k becomes the regularization parameter and the method is called Truncated SVD.

Another possible scheme is

$$x_\alpha = \operatorname*{arginf}_{x \in \mathcal{X}} \left\{ \|y - Tx\| + \alpha^2 \|x\|^2 \right\}.$$

This is called Tikhonov regularization and the regularization parameter clearly controls the balance between minimizing the norm of the residual and the size of the solution. The regularized inverse $T_\alpha^\#$ can be expressed explicitly by:

$$T_\alpha^\# = (\alpha^2 I + T^* T)^{-1} T^*.$$

A thorough survey of regularization schemes and ways to find the optimal regularization parameter can be found in [Han98a].

[3]This fact in itself is not necessarily a bad thing since one can easily have pointwise convergence without operator norm convergence.

5.4 Iterative Methods

Another way of regularizing is by using iterative methods. Here the hope is that the iteration number *itself* acts as a regularization parameter. Before digging more into this, let us introduce some key quantities and notation.

Let the exact solution be $x^\star$ and let $x^{(0)}$ denote a starting guess (often the zero vector). Similarly we let $x^{(k)}$, $k = 1, 2, \ldots$, denote the solution found by the iterative method after iteration k.

We furthermore define the important quantities

$$\text{Residual:} \quad r(x) = b - Ax \quad \text{and} \quad r^{(k)} = r(x^{(k)}) \tag{5.7}$$

$$\text{Error:} \quad e(x) = x^\star - x \quad \text{and} \quad e^{(k)} = e(x^{(k)}) \tag{5.8}$$

for each iteration $k \ge 0$. Notice that only the residual and not the error will be computable by an iterative method. Important to note is then

$$Ae^{(k)} = Ax^\star - Ax^{(k)} = b - Ax^{(k)} = r^{(k)}.$$

Let us consider this equality when A is represented by a (finite) square matrix. If A does not have an inverse then A has a non-trivial null-space and $r^{(k)} = 0$ will *not* imply $e^{(k)} = 0$. If A does have an inverse we get

$$\|r^{(k)}\|_2 = \|A^{-1} e^{(k)}\|_2 \le \|A^{-1}\|_2 \|e^{(k)}\|_2.$$

This means that $\|e^{(k)}\|_2 \to 0$ implies $\|r^{(k)}\|_2 \to 0$, but since $\|A^{-1}\|_2$ is typically very large for ill-posed problems (reflecting the unboundedness of the inverse of a compact operator), we must be careful not to deduce that the error norm is small just because the residual norm is.

A good example of this is shown in Section D.3 in the appendix, where the iterative method GMRES is used. For this particular example, the residual norm decreases fast while the error norm remains almost constant.

5.4.1 Rate of Convergence

Consider the quantity

$$\gamma_k = \left(\frac{\|e^{(k)}\|_2}{\|e^{(0)}\|_2} \right)^{\frac{1}{k}}, \tag{5.9}$$

which can be considered the average reduction factor of the error norm per iteration. Using this, we define

$$R_k = -\log_{10}(\gamma_k)$$

as the *average rate of convergence*. In some cases, this quantity is closely related to the minimum number of iterations required to obtain a certain accuracy (see also [You71]).

Important to note is that sometimes, as is the case in this thesis, it is only possible to estimate γ_k from Equation (5.9) using the *residual* norms and not the *error* norms. As previously mentioned, this will not always reveal the true behavior of the method.

5.4.2 Semiconvergence

Let us now return to the subject of using iterative methods as regularization schemes. As noted earlier, when dealing with a discrete ill-posed problem and a noisy right-hand side b_δ, we are not interested in the exact solution $x^\star_\delta = A^{-1}b_\delta$. Instead, we are interested in the solution computed in infinite precision from the noiseless right-hand side $x^\star = A^{-1}b$. Obtaining this solution is of course a little too optimistic, but we can hope to come close at some iteration step.

Consider the plot in Figure 5.2. Here, the residual and error norms are shown for each iteration. Note however, that it is not the error compared to the true solution $x^\star_\delta$, but compared to the solution computed from a noiseless right-hand side $x^\star$. Although the residual norm continuously decreases, the shown error norm first decreases, but then starts to *diverge* from the wanted solution. This can often be explained by the fact that the wanted solution typically is very smooth, and that

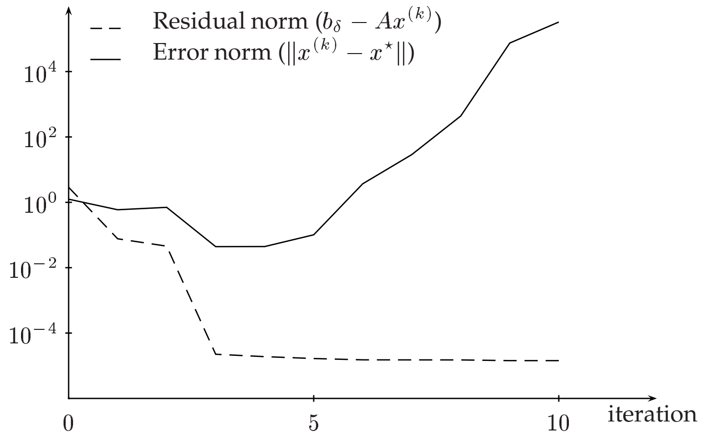
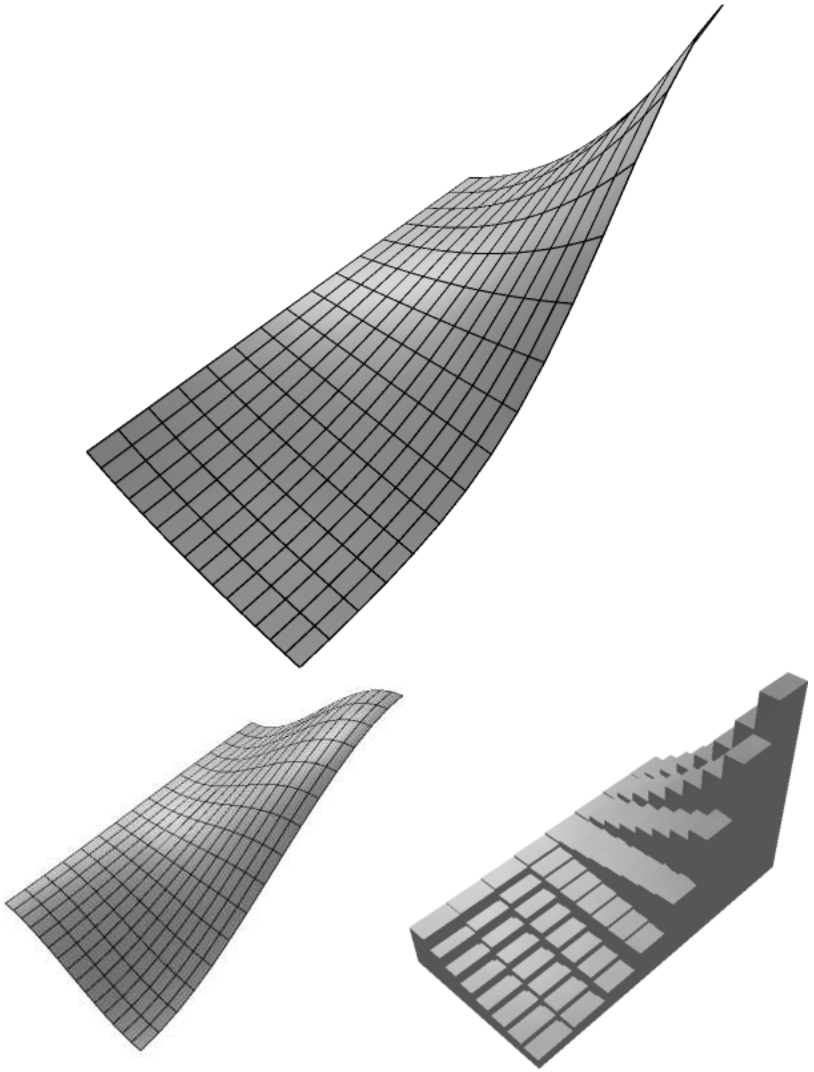


Figure 5.2: *An illustration of semiconvergence. Note that the error is calculated using the true, noiseless solution.*

the smooth components of the solution often are obtained in the first iteration steps. In later steps, the noise of the right-hand side begins to influence the solution and the divergence sets in. This phenomenon is often termed *semiconvergence*.[4]

[4]The term semiconvergence is used by [Han98a], who has adopted it from F. Natterer.



CHAPTER 6

Approximating in Finite Dimensions

approximate, *to come close to; be nearly the same as*
— THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE

finite, *limited in quantity, degree, or capacity*
— WEBSTER'S REVISED UNABRIDGED DICTIONARY

What happens to the behavior of operators, eigenvalues, singular values and such when approximating an operator by a finite dimensional mapping? This is a very important topic, since all infinite problems that can not be solved analytically, need to be solved approximately on computer. And since a computer can not represent infinite amounts of data, we have to settle with finite dimension computations.

6.1 The Galerkin Method

Consider the equation $Kf = g$ where $K \in B(\mathcal{X}, \mathcal{Y})$ is a compact operator. Let $\mathcal{X}_N \subset \mathcal{X}$ and $\mathcal{Y}_M \subset \mathcal{Y}$ be subspaces with dimensions N and

M respectively. If a function $f_N \in \mathcal{X}_N$ satisfies

$$(Kf_N, y) = (g, y) \quad \text{for all } y \in \mathcal{Y}_M, \tag{6.1}$$

we call f_N a solution to the *Galerkin equation*.

Let now (ϕ_j) be an orthonormal basis for $\mathcal{X}_N$ and let (ψ_j) be an orthonormal basis for $\mathcal{Y}_M$. Then we can write any $f_N \in \mathcal{X}_N$ as $f_N = \sum_{j=1}^N \mathbf{x}_j \phi_j$ and if we set $\mathbf{y}_i = (g, \psi_i)$ the Galerkin Equation (6.1) becomes

$$\begin{aligned} \left(K\left(\sum_{j=1}^N \mathbf{x}_j\phi_j\right), \psi_i\right) &= (g, \psi_i) \quad \Leftrightarrow \\ \sum_{j=1}^N \mathbf{x}_j (K\phi_j, \psi_i) &= \mathbf{y}_i \quad \text{for all } i = 1, 2, \ldots, M. \end{aligned}$$

By introducing a matrix $\mathbf{A} \in \mathbb{C}^{M\times N}$ with entries $\mathbf{A}_{i,j} = (K\phi_j, \psi_i)$ we see that this corresponds to a simple matrix equation

$$\mathbf{A}\mathbf{x} = \mathbf{y}. \tag{6.2}$$

So given an equation of the form $Kf = g$, an approximate solution can be found by computing $\mathbf{y}_i = (g, \psi_i)$ and $\mathbf{A}_{i,j} = (K\phi_j, \psi_i)$, solving Equation (6.2) for $\mathbf{x}$ and finally computing $\widetilde{f} = \sum_{j=1}^N \mathbf{x}_j\phi_j$. Note that this requires computing:

- One integral for each i to compute $\mathbf{y}$.
- Two integrals for each i and j to compute $\mathbf{A}$ (if K is an integral operator).

Most often in practise, these integrals have to be approximated using e.g. collocation or quadrature methods, see [Bak77] or others. Note that these approximation and round-off errors connected herewith *will be ignored* in this thesis.

An example of discretizing an operator $K : L^2(I \times I) \to L^2(I \times I)$, represented by a *double* integral, can be seen in Section D.5 in the appendix. Here, a simple quadrature formula is used to approximate an integral.

It is easy to see that the matrix $\mathbf{A}$ represents the mapping $\widetilde{K}_{MN} : \mathcal{X} \to \mathcal{Y}$ defined as

$$\widetilde{K}_{MN}x = \Pi_{\mathcal{Y}_M} K \Pi_{\mathcal{X}_N} x = \sum_{i=1}^M \sum_{j=1}^N (K\phi_j, \psi_i)(x, \phi_j)\psi_i, \tag{6.3}$$

where $\Pi_{\mathcal{X}_N}$ and $\Pi_{\mathcal{Y}_M}$ are orthogonal projections onto $\mathcal{X}_N$ and $\mathcal{Y}_M$ respectively (see Equation (2.11) on page 18).

We will now show that $\widetilde{K}_{MN} \to K$ in the operator norm as $M, N \to \infty$. First, we need some preliminaries. Let $L \in B(\mathcal{X}, \mathcal{Y})$ be an arbitrary compact operator and let $\Pi_N : \mathcal{Y} \to \mathcal{Y}$ be a projection onto an N-dimensional subspace. One can show that

$$\sup_{x\in M} \|\Pi_N x - x\| \to 0, \tag{6.4}$$

for any subset $M \subset \mathcal{Y}$ for which $\overline{M}$ is compact (see Lemma 4.3.7 in [Hac95]). Consider now the set $B = \{Lx \mid \|x\| \le 1\}$. Since L is compact, $\overline{B}$ is compact by definition. The above result now yields

$$\|\Pi_N L - L\| = \sup_{\|x\|\le 1} \|\Pi_N Lx - Lx\| = \sup_{y\in B} \|\Pi_N y - y\| \to 0,$$

as $N \to \infty$. So $\Pi_N L \to L$ in the operator norm. We also have $L\Pi_N \to L$. This is seen by

$$L\Pi_N = ((L\Pi_N)^*)^* = (\Pi_N L^*)^* \to (L^*)^* = L$$

where we have used that $\Pi_N^* = \Pi_N$ and that the adjoint of a compact operator also is compact.

The convergence of $\widetilde{K}_{MN}$ to K can now be seen by

$$\begin{aligned} \|\widetilde{K}_{MN} - K\| &= \|\Pi_{\mathcal{Y}_M} K \Pi_{\mathcal{X}_N} - K\| \\ &\le \|\Pi_{\mathcal{Y}_M} K \Pi_{\mathcal{X}_N} - \Pi_{\mathcal{Y}_M} K\| + \|\Pi_{\mathcal{Y}_M} K - K\| \to 0. \end{aligned}$$

Since a projection is bounded, $\Pi_{\mathcal{Y}_M} K$ is a compact operator which means that $(\Pi_{\mathcal{Y}_M} K)\Pi_{\mathcal{X}_N} \to \Pi_{\mathcal{Y}_M} K$ from the results just shown above. Hence, the above limit.

So $\widetilde{K}_{MN}$ converges to K in the operator norm. This is a very strong result which of course implies pointwise convergence. That a discretization scheme converges pointwise is sometimes called *consistent*.

The choice of basis and how quickly $\widetilde{K}_{MN}$ actually converges to K is obviously of great importance and relevance. No results in this thesis will be related to this, however. A lot of literature covers this and it was decided, in relation to this thesis, that other areas were of greater importance. See [Bak77], [Hac95], [Kre99] or others for results on operator approximations.

We shall in a short while use the concept *collectively compact*. This is defined as follows.

Definition 6.1 *A set $\{T_n\}_{n\in\mathbb{N}}$ of operators $T_n \in B(\mathcal{X},\mathcal{Y})$ is called* ***collectively compact****, if the set*

$$\overline{\{T_n x \mid x \in \mathcal{X}, \|x\| \leq 1, n \in \mathbb{N}\}}$$

is compact.

It can be shown that if each operator T_n is compact and T_n converges in the operator norm, then $\{T_n\}$ is collectively compact (page 134 in [Hac95]). Since each $\widetilde{K}_{MN}$ is finite dimensional it is compact, and as just seen, $\widetilde{K}_{MN}$ converges in norm to K. So $\{\widetilde{K}_{MN}\}$ is collectively compact. Later, it will become clear why this property is important.

We now return to look at possible relations between $\widetilde{K}_{MN}$ and its matrix representation $\mathbf{A}$. The singular values of $\mathbf{A}$ are identical to those of $\widetilde{K}_{MN}$: Assume that $\mathbf{A}\mathbf{v} = \mu\mathbf{u}$ and consider

$$\begin{aligned}
\mathbf{A}\mathbf{v} = \mu\mathbf{u} \quad &\Leftrightarrow \quad \sum_{j=1}^{N}(K\phi_j,\psi_i)(v,\phi_j) = \mu(u,\psi_i), \quad i = 1,\ldots,M \quad \Leftrightarrow \\
\sum_{i=1}^{M}\sum_{j=1}^{N}(K\phi_j,\psi_i)(v,\phi_j)\psi_i &= \mu\sum_{i=1}^{M}(u,\psi_i)\psi_i \quad \Leftrightarrow \quad \widetilde{K}_{MN}v = \mu u,
\end{aligned} \tag{6.5}$$

where

$$v = \sum_{j=1}^{N}\mathbf{v}_j\phi_j \quad \text{and} \quad u = \sum_{i=1}^{M}\mathbf{u}_i\psi_i.$$

These expression show at the same time the correspondence between the singular vectors of $\mathbf{A}$ and $\widetilde{K}_{MN}$.

What about eigenvalues and -vectors? If the eigenvalue problem has to make sense for $\mathbf{A}$ we must have $M = N$. And if eigenvectors of $\mathbf{A}$ has to correspond to eigenvectors of $\widetilde{K}_{MN}$ we must have $\mathcal{X} = \mathcal{Y}$ and that the bases used in $\mathcal{X}$ and $\mathcal{Y}$ are identical (see also Section 4.3.3). So when assuming $M = N$ and $\psi_i = \phi_i$ for $i = 1, 2, \ldots, N$ we have, analogous to Equation (6.5):

$$\begin{aligned}
\mathbf{A}\boldsymbol{\varphi} = \lambda\boldsymbol{\varphi} \quad &\Leftrightarrow \quad \widetilde{K}_{MN}\varphi = \lambda\varphi, \\
\text{where} \quad &\varphi = \sum_{i=1}^{N}\boldsymbol{\varphi}_i\phi_i.
\end{aligned}$$

Since we are converting to the world of (finite) matrices, we will now formally introduce the adjoint of a matrix mapping. Consider an operator $T : \mathbb{C}^N \to \mathbb{C}^N$ represented by a matrix $\mathbf{T}$. For $x, y \in \mathbb{C}^N$, represented by the vectors $\mathbf{x}$ and $\mathbf{y}$ respectively, we have

$$\begin{aligned}
(Tx,y) &= \sum_{i=1}^{N}(Tx)_i\overline{\mathbf{y}_i} = \sum_{i=1}^{N}\left(\sum_{j=1}^{N}\mathbf{T}_{i,j}\mathbf{x}_j\right)\overline{\mathbf{y}_i} = \sum_{j=1}^{N}\mathbf{x}_j\left(\sum_{i=1}^{N}\mathbf{T}_{i,j}\overline{\mathbf{y}_i}\right) \\
&= \sum_{i=1}^{N}\mathbf{x}_i\overline{\sum_{j=1}^{N}\overline{\mathbf{T}_{j,i}}\mathbf{y}_j} = (x, T^*y),
\end{aligned}$$

which means that $(T^*y)_i = \sum_{j=1}^{N}\overline{\mathbf{T}_{j,i}}\mathbf{y}_j$. From this follows that the adjoint operator T^* is represented by a matrix with entries $(\overline{\mathbf{T}_{j,i}})$, the conjugate transpose or Hermitian of $\mathbf{T}$, written $\mathbf{T}^H$.

6.2 Singular Value Decomposition

Let N' denote the number of non-zero singular values of $\widetilde{K}_{MN}$, $N' \leq \min(M,N)$. This means we have

$$\widetilde{K}_{MN} f = \sum_{n=1}^{N'} \sigma_n (f, \tilde{v}_n) \tilde{u}_n.$$

Let $\mathbf{A}$ be the matrix representing this mapping with respect to the bases $(\phi_n) \subset \mathcal{X}$ and $(\psi_n) \subset \mathcal{Y}$. Let the matrices $\mathbf{V} \in \mathbb{C}^{N \times N'}$ and $\mathbf{U} \in \mathbb{C}^{M \times N'}$ be defined as

$$\mathbf{V}_{i,j} = (\tilde{v}_j, \phi_i) \quad \text{and} \quad \mathbf{U}_{i,j} = (\tilde{u}_j, \psi_i),$$

for all relevant indices. They now represent the singular vectors with respect to the chosen bases. Consider an entry of $\mathbf{A}$:

$$\begin{aligned} \mathbf{A}_{i,j} = (K\phi_j, \psi_i) &= \left(\sum_{n=1}^{N'} \sigma_n (\phi_j, \tilde{v}_n) \tilde{u}_n, \psi_i \right) = \sum_{n=1}^{N'} \sigma_n (\phi_j, \tilde{v}_n)(\tilde{u}_n, \psi_i) \\ &= \sum_{n=1}^{N'} \sigma_n \overline{\mathbf{V}_{j,n}} \mathbf{U}_{i,n} = \sum_{n=1}^{N'} \mathbf{U}_{i,n} \sigma_n \left(\mathbf{V}^H\right)_{n,j} = \left(\mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H\right)_{i,j} \end{aligned} \tag{6.6}$$

where $\boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \sigma_2, \ldots, \sigma_{N'})$. So the singular value decomposition for matrices becomes

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^H.$$

Results will now be derived that show how well singular values and singular vectors are approximated by a finite dimensional operator. The results are almost identical to those in [Han88] but the presentation is somewhat different.

The notation $\widetilde{K}_N$ will be used for an approximating operator defined as in Equation 6.3 with $M = N$. Introducing the N-dimensional spaces $\mathcal{X}_N = \operatorname{span}\{\phi_1, \phi_2, \ldots, \phi_N\}$ and $\mathcal{Y}_N = \operatorname{span}\{\psi_1, \psi_2, \ldots, \psi_N\}$ we have $\widetilde{K}_N : \mathcal{X}_N \to \mathcal{Y}_N$. We will use the notation σ_i^N for the ith singular vector of the operator $\widetilde{K}_N$.

We can now state the following very important theorems.

Theorem 6.2 *The singular values of an approximating operator $\widetilde{K}_N$ approaches the singular values of K from* below. *More precisely:*

$$\sigma_i^N \leq \sigma_i^{N+1} \leq \mu_i, \quad \textit{for all } i = 1, 2, \ldots, N. \tag{6.7}$$

Proof: From the expressions in (3.10) we have:

$$\begin{aligned} \sigma_1^N &= \sup_{\substack{x \in \mathcal{X}_N \\ \|x\|=1}} \|\widetilde{K}_N x\| \quad \text{and} \\ \sigma_{i+1}^N &= \inf_{z_1, z_2, \ldots, z_i \in \mathcal{X}_N} \sup_{\substack{x \in \mathcal{X}_N \\ \|x\|=1 \\ x \perp z_1, z_2, \ldots, z_i}} \|\widetilde{K}_N x\|. \end{aligned} \tag{6.8}$$

Since we always have $x \in \mathcal{X}_N$ we get

$$\begin{aligned} \|\widetilde{K}_N x\|^2 = (\widetilde{K}_N x, \widetilde{K}_N x) &= \sup_{y \in \mathcal{Y}_N} |(\widetilde{K}_N x, y)| \\ &= \sup_{y \in \mathcal{Y}_N} |(Kx, y)| = \left(\sup_{y \in \mathcal{Y}_N, \|y\|=1} |(Kx, y)| \right)^2. \end{aligned}$$

The second equality sign stems from the fact that $\widetilde{K}_N \mathcal{X}_N \subset \mathcal{Y}_N$ and that for fixed x, the quantity $|(\widetilde{K}_N x, y)|$ attains its maximum exactly when $y = \widetilde{K}_N x$.

So now the expressions in (6.8) can be rewritten to

$$\begin{aligned} \sigma_1^N &= \sup_{\substack{x \in \mathcal{X}_N,\, y \in \mathcal{Y}_N \\ \|x\|=1, \|y\|=1}} |(Kx, y)| \quad \text{and} \\ \sigma_{i+1}^N &= \inf_{z_1, z_2, \ldots, z_i \in \mathcal{X}_N} \sup_{\substack{x \in \mathcal{X}_N,\, y \in \mathcal{Y}_N \\ x \perp z_1, z_2, \ldots, z_i \\ \|x\|=1, \|y\|=1}} |(Kx, y)|. \end{aligned}$$

Now consider these expressions for σ_i^N and σ_i^{N+1}. These are identical except that the sets from which the x's and y's are chosen, are *larger* in the expression for σ_i^{N+1}. Hence, $\sigma_i^N \le \sigma_i^{N+1}$.

When letting $N \to \infty$ we get $\widetilde{K}_N \to K$ and this leads to $\sigma_i^{N+1} \le \sigma_i^{N+2} \le \cdots \le \mu_i$. □

The quantity δ_N, defined as

$$\delta_N^2 = \|K\|_2^2 - \|\widetilde{K}_N\|_2^2 = \sum_{i=1}^{\infty} \mu_i^2 - \sum_{i=1}^{N} \left(\sigma_i^N\right)^2,$$

turns out to be important in the following estimations. Note that the above theorem ensures that the right-hand side is positive.

The next theorem shows how much the *total* difference between the true and approximate singular values can be.

Theorem 6.3 *The errors of the approximate singular values fulfill the inequality*

$$\sum_{i=1}^{N} \left(\mu_i - \sigma_i^N\right)^2 \le \delta_N^2. \tag{6.9}$$

Proof: By using $\mu_i \le \sigma_i^N$ we get by straightforward calculation:

$$\begin{aligned}
\sum_{i=1}^{N} \left(\mu_i - \sigma_i^N\right)^2 &= \sum_{i=1}^{N} \mu_i^2 + \sum_{i=1}^{N} \left(\sigma_i^N\right)^2 - 2\sum_{i=1}^{N} \mu_i \sigma_i^N \\
&\le \sum_{i=1}^{N} \mu_i^2 + \sum_{i=1}^{N} \left(\sigma_i^N\right)^2 - 2\sum_{i=1}^{N} \left(\sigma_i^N\right)^2 \\
&\le \sum_{i=1}^{\infty} \mu_i^2 - \sum_{i=1}^{N} \left(\sigma_i^N\right)^2 = \delta_N^2.
\end{aligned}$$

□

An approximate singular value lies in a small interval below the true singular value, more precisely:

Theorem 6.4 *The approximate singular values are bounded by the true ones in the following way:*

$$\mu_i^2 - \delta_N^2 \le \left(\sigma_i^N\right)^2 \le \mu_i^2, \quad \textit{for all } i = 1, 2, \ldots, N. \tag{6.10}$$

Proof: The inequalities in (6.7) says that $\sigma_i^N \le \mu_i$ for all $i = 1, 2, \ldots, N$. This leads to the rightmost inequality. It also leads, for a fixed i, to:

$$\begin{aligned}
\sum_{k=1}^{N} \left(\sigma_k^N\right)^2 - \left(\sigma_i^N\right)^2 &\le \sum_{k=1}^{N} \mu_k^2 - \mu_i^2 \quad \Leftrightarrow \\
\mu_i^2 - \left(\sigma_i^N\right)^2 &\le \sum_{k=1}^{N} \mu_k^2 - \sum_{k=1}^{N} \left(\sigma_k^N\right)^2 \\
&\le \sum_{k=1}^{\infty} \mu_k^2 - \sum_{k=1}^{N} \left(\sigma_k^N\right)^2 = \delta_N^2.
\end{aligned}$$

□

We will in the following leave out the superscript on the approximate singular values and just use σ_i since only a fixed value of N is considered.

Theorem 6.5 *When a smallest integer m exists such that $i \le m < N$ and $\mu_{m+1} < \mu_i$ then the error of the approximate singular vectors are bounded by*

$$\max\left\{\|\tilde{v}_i - v_i\|, \|\tilde{u}_i - u_i\|\right\} \le \sqrt{2\frac{\mu_i - \sigma_i + \displaystyle\sum_{k=1,k\ne i}^{m} \mu_k |(\tilde{u}_i, u_k)||(\tilde{v}_i, v_k)|}{\mu_i - \mu_{i+1}}}. \tag{6.11}$$

Proof: Assume there exists a smallest integer $i \le m < N$ such that $\mu_{m+1} < \mu_i$. Let $\alpha_k = (\tilde{u}_i, u_k)$ and $\gamma_k = (\tilde{v}_i, u_k)$ for $k = 1, 2, \ldots, m$. Define the *residual* vectors

$$u_0 = \tilde{u}_i - \sum_{k=1}^{m} \alpha_k u_k, \qquad v_0 = \tilde{v}_i - \sum_{k=1}^{m} \gamma_k v_k.$$

The norms of these are

$$\|u_0\|^2 = (u_0, u_0) = 1 - \sum_{k=1}^{m} |\alpha_k|^2, \quad \text{and} \quad \|v_0\|^2 = 1 - \sum_{k=1}^{m} |\gamma_k|^2.$$

We now get

$$\begin{aligned}(Kv_0, u_0) = (K\tilde{v}_i, \tilde{u}_i) - \sum_{k=1}^{m} \alpha_k (K\tilde{v}_i, u_k) - \sum_{k=1}^{m} \gamma_k (Kv_k, \tilde{u}_i) \\ + \sum_{j=1}^{m}\sum_{k=1}^{m} \gamma_j \alpha_k (Kv_j, u_k) = \sigma_i - \sum_{k=1}^{m} \mu_k \overline{\alpha_k} \gamma_k.\end{aligned} \tag{6.12}$$

Here it has been used that $(K\tilde{v}_i, u_k) = (\tilde{v}_i, K^* u_k) = (\tilde{v}_i, \mu_k v_k) = \mu_k \gamma_k$.

Since u_0 by construction is perpendicular to $u_1, u_2, \ldots, u_m$ and analogous for v_0 we have

$$|(Kv_0, u_0)| \leq \mu_{m+1} \|v_0\| \|u_0\|.$$

By inserting the expression (6.12) on the left hand side we get

$$\begin{aligned}\left|\sigma_i - \sum_{k=1}^{m} \mu_k \overline{\alpha_k} \gamma_k\right| &\leq \mu_{m+1} \sqrt{1 - \sum_{k=1}^{m} |\alpha_k|^2} \sqrt{1 - \sum_{k=1}^{m} |\gamma_k|^2} \quad \Leftrightarrow \\ \sigma_i - \sum_{k=1}^{m} \mu_k |\alpha_k| |\gamma_k| &\leq \mu_{m+1} (1 - |\alpha_k| |\gamma_k|\end{aligned}$$

since $(1-x^2)(1-y^2) \leq (1-xy)^2$ for all $x, y \in \mathbb{R}$.

By appropriately rearranging we obtain

$$1 - |\alpha_i| |\gamma_i| \leq \frac{\mu_i - \sigma_i + \sum_{k=1, k \neq i}^{m} \mu_k |\alpha_k| |\gamma_k|}{\mu_i - \mu_{i+1}}. \tag{6.13}$$

We now seek to express the errors of the singular vectors in terms of the quantities found above. We find

$$\|\tilde{u}_i - u_i\|^2 = 2 - (\overline{\alpha_i} + \alpha_i) \leq 2(1 - |\alpha_i|) \text{ and } \|\tilde{v}_i - v_i\|^2 \leq 2(1 - |\gamma_i|).$$

Using this we get

$$\max\left\{\|\tilde{v}_i - v_i\|^2, \|\tilde{u}_i - u_i\|^2\right\} \leq 2 \max\{1 - |\alpha_i|, 1 - |\gamma_i|\} \leq 2(1 - |\alpha_i||\gamma_i|)$$

which leads to the desired result by inserting (6.13). □

By using the bounds $\mu_i - \sigma_i^N \leq \delta_N$ and

$$|(\tilde{v}_i, v_k)| = |(\tilde{v}_i, \tilde{v}_k + v_k - \tilde{v}_k)| = |(\tilde{v}_i, \tilde{v}_k)| + |(\tilde{v}_i, v_k - \tilde{v}_k)| \leq \|v_k - \tilde{v}_k\|,$$

and similarly for $|(\tilde{u}_i, u_k)|$, the bound in (6.11) becomes looser but a bit more practically useful.

6.3 Eigenvalue Bounds

We will now look at approximation errors of eigenvalues and -vectors. The results are mainly from [Hac95], but some similar results also appear in [Ans71].

Theorem 6.6 *Let $(\widetilde{K}_n)$ be collectively compact and consistent with the compact operator $K \in B(\mathcal{X})$. Let furthermore (λ_n) be a sequence of eigenvalues fulfilling $\widetilde{K}_n \varphi_n = \lambda_n \varphi_n$ and $\|\varphi_n\| = 1$ for all $n \in \mathbb{N}$.*

Then there exists a subsequence (λ_{n_k}) either converging to zero or to an eigenvalue λ of K. If $\lambda = \lim_{k\to\infty} \lambda_{n_k}$ is non-zero, the subsequence can be chosen so that (φ_{n_k}) converge to an eigenfunction φ of K corresponding to λ.

Proof: Consider the set $B = \{\widetilde{K}_n \varphi_n \mid n \in \mathbb{N}\}$. Because of the collective compactness and since $\|\varphi_n\| = 1$, we have that $\overline{B}$ is compact so $C = \sup\{\|x\| \mid x \in B\}$ exists. We now have $|\lambda_n| = |\lambda_n| \|\varphi_n\| = \|\widetilde{K}_n \varphi_n\| \leq C$

which shows that $\{\lambda_n\}$ lies in a compact set. Since $\overline{B}$ was compact, we can choose indices such that both

$$\lambda_{n_i} \to \lambda \quad \text{and} \quad \widetilde{K}_{n_i}\varphi_{n_i} \to x$$

are fulfilled.

Assume that $\lambda \neq 0$. Then $\varphi_{n_i} = \lambda_{n_i}^{-1}\widetilde{K}_n\varphi_n$ converges to $\varphi = \lambda^{-1}x$. Consider now

$$\widetilde{K}_{n_i}\varphi_{n_i} = \widetilde{K}_{n_i}(\varphi_{n_i} - \varphi) + (\widetilde{K}_{n_i} - K)\varphi + K\varphi.$$

The first term on the right-hand side tends to zero because of the continuity of $\widetilde{K}_{n_i}$. The second term tends to zero because of pointwise convergence (consistency). So we have $\widetilde{K}_{n_i}\varphi_{n_i} \to K\varphi$. Since $\|\varphi\| = \lim_{i\to\infty}\|\varphi_{n_i}\| = 1$ we see that $\varphi \neq 0$ and since $\widetilde{K}_{n_i}\varphi_{n_i} = \lambda_{n_i}\varphi_{n_i} \to \lambda\varphi$ we have $K\varphi = \lambda\varphi$. □

So given a sequence of eigenvalues (λ_n), each corresponding to the approximate operator $\widetilde{K}_n$, and λ_n converges to a non-null element λ, then λ will be an eigenvalue of K. In Section D.1 in the appendix, there is an example of a discretization where every eigenvalue of $\widetilde{K}_n$ converges to zero as $n \to \infty$.

As the following theorem will show, the converse of Theorem 6.6 is also true.

Theorem 6.7 *Let $(\widetilde{K}_n)$ be a sequence of compact operators, consistent with the compact operator $K \in B(\mathcal{X})$. Then for any eigenvalue $\lambda \neq 0$ of K there exists is a sequence (λ_n), where λ_n is an eigenvalue of K_n for all n, such that $\lambda_n \to \lambda$.*

Proof: See Theorem 4.8.16 in [Hac95]. □

We now turn to look at how much the approximate eigenvalues and -vectors can differ from the true ones.

Theorem 6.8 *Let $(\widetilde{K}_n)$ be a sequence of compact operators, consistent with the compact operator $K \in B(\mathcal{X})$ and assume*

$$\|K - \widetilde{K}_n\| = \mathcal{O}(n^{-p}), \quad p > 1.$$

Let $\lambda \neq 0$ be a simple eigenvalue of K with corresponding eigenfunction φ and let φ^* satisfy $K^*\varphi^* = \lambda\varphi^*$. The dual operators $\widetilde{K}_n^*$ must satisfy $\widetilde{K}_n^*\varphi^* \to K^*\varphi^*$. Let (λ_n) be a sequence of eigenvalues satisfying $\lambda_n \to \lambda$ and $\widetilde{K}_n\varphi_n = \lambda_n\varphi_n$ for each n. Let furthermore the eigenfunctions be scaled by $(\varphi^*, \varphi) = 1$ and $(\varphi^*, \varphi_n) = 1$.

Then the following asymptotic bounds hold:

$$|\lambda - \lambda_n| = \mathcal{O}(n^{-p}) \quad \textit{and} \quad \|\varphi - \varphi_n\| = \mathcal{O}(n^{-p}). \tag{6.14}$$

Proof: Observe the following identity

$$\lambda_n(\varphi_n - \varphi) = \widetilde{K}_n(\varphi_n - \varphi) + d_n \tag{6.15}$$

where

$$d_n = \lambda^{-1}\left[\lambda_n(\widetilde{K}_n - K)\varphi + (\lambda - \lambda_n)\widetilde{K}_n\varphi\right].$$

This last quantity can be bounded by

$$\begin{aligned}\|d_n\| &\leq |\lambda|^{-1}\left[|\lambda_n|\|\widetilde{K}_n - K\|\|\varphi\| + |\lambda - \lambda_n|\|\widetilde{K}_n\varphi\|\right] \\ &\leq \mathcal{O}(n^{-p}) + c_1|\lambda - \lambda_n|.\end{aligned}$$

Because of Equation (6.15) and the assumption $(\varphi^*, \varphi) = (\varphi^*, \varphi_n) = 1$ we see that the matrix equation

$$\begin{bmatrix}\lambda_n I - \widetilde{K}_n & \varphi \\ \varphi^* & 0\end{bmatrix}\begin{bmatrix}\varphi_n - \varphi \\ 0\end{bmatrix} = \begin{bmatrix}d_n \\ 0\end{bmatrix}$$

is fulfilled. This means that given the above right-hand side, the equation can be solved. But is the inverse of the operator matrix on the left-hand side bounded? Yes it is, since the inverse can be written explicitly:

$$\begin{bmatrix}\lambda_n I - \widetilde{K}_n & \varphi \\ \varphi^* & 0\end{bmatrix}\begin{bmatrix}0 & \varphi \\ \varphi^* & \widetilde{K}_n - \lambda_n I\end{bmatrix} = \begin{bmatrix}I & 0 \\ 0 & 1\end{bmatrix}$$

and similarly if multiplied from the right.

This yields

$$\|\varphi_n - \varphi\| = \left\| \begin{bmatrix} \varphi_n - \varphi \\ 0 \end{bmatrix} \right\| \leq \left\| \begin{bmatrix} 0 & \varphi \\ \varphi^* & \widetilde{K}_n - \lambda_n I \end{bmatrix} \right\| \left\| \begin{bmatrix} d_n \\ 0 \end{bmatrix} \right\| \qquad (6.16)$$
$$= c_2\|d_n\| = \mathcal{O}(n^{-p}) + c_3|\lambda - \lambda_n|.$$

Consider now the identity

$$(\lambda I - K)(\varphi - \varphi_n) = (\lambda_n - \lambda)\varphi_n - (\widetilde{K}_n - K)\varphi - (\widetilde{K}_n - K)(\varphi_n - \varphi).$$

We now wish apply the functional $(\varphi^*, \cdot)$ to each side of this expression. The left-hand side becomes

$$(\varphi^*, (\lambda I - K)(\varphi - \varphi_n)) = ((\lambda I - K^*)\varphi^*, \varphi - \varphi_n) = 0.$$

The right-hand side:

$$\Big(\varphi^*, (\lambda_n - \lambda)\varphi_n - (\widetilde{K}_n - K)\varphi - (\widetilde{K}_n - K)(\varphi_n - \varphi)\Big)$$
$$= \overline{\lambda_n - \lambda} - \Big(\varphi^*, (\widetilde{K}_n - K)\varphi - (\widetilde{K}_n - K)(\varphi_n - \varphi)\Big).$$

Combining the two sides and by using the Cauchy-Schwartz inequality we get

$$\begin{aligned} |\lambda_n - \lambda| &= \Big|\Big(\varphi^*, (\widetilde{K}_n - K)\varphi - (\widetilde{K}_n - K)(\varphi_n - \varphi)\Big)\Big| \\ &\leq \|\varphi^*\| \Big\|(\widetilde{K}_n - K)\varphi - (\widetilde{K}_n - K)(\varphi_n - \varphi)\Big\| \\ &= c_4\left(\mathcal{O}(n^{-p})c_5 + \mathcal{O}(n^{-p})\|\varphi_n - \varphi\|\right). \end{aligned}$$

Since $\varphi_n \to \varphi$ we have $\|\varphi_n - \varphi\| \leq c_6$ for all $n \in \mathbb{N}$ for some $c_6 \in \mathbb{R}$ so $|\lambda_n - \lambda| = \mathcal{O}(n^{-p})$. Inserting this bound into Equation (6.16) gives us $|\varphi_n - \varphi| = \mathcal{O}(n^{-p})$. □

When Galerkin discretization is used, Hackbusch [Hac95] has shown that the eigenvalues actually converge as $|\lambda - \lambda_n| = \mathcal{O}(n^{-2p})$.

6.4 Relating Singular Values and Eigenvalues

The results for relating singular values and eigenvalues for general compact operators obviously still hold in finite dimensions (see Section 3.3). But can more be said? Yes, a little. Assume that $K \in B(\mathcal{X})$ and that $\mathcal{X}$ is of dimension N. Assume furthermore that K is has no zero eigenvalues, or equivalently, no zero singular values (K and K^*K have identical null-spaces). Then K can be represented both as

$$Kx = \sum_{n=1}^{N} \lambda_n (x, \varphi_n)\varphi_n \quad \text{and} \quad Kx = \sum_{n=1}^{N} \mu_n (x, v_n) u_n.$$

From Theorem 3.7 we have the following inequality,

$$\prod_{n=1}^{N} |\lambda_n| \leq \prod_{n=1}^{N} \mu_n. \qquad (6.17)$$

Since $\lambda_n \neq 0$ and $\mu_n \neq 0$ for all $n = 1, 2, \ldots, N$, the inverse K^{-1} exists with the obvious representations

$$K^{-1}x = \sum_{n=1}^{N} \frac{1}{\lambda_n}(x, \varphi_n)\varphi_n \quad \text{and} \quad K^{-1}x = \sum_{n=1}^{N} \frac{1}{\mu_n}(x, u_n) v_n.$$

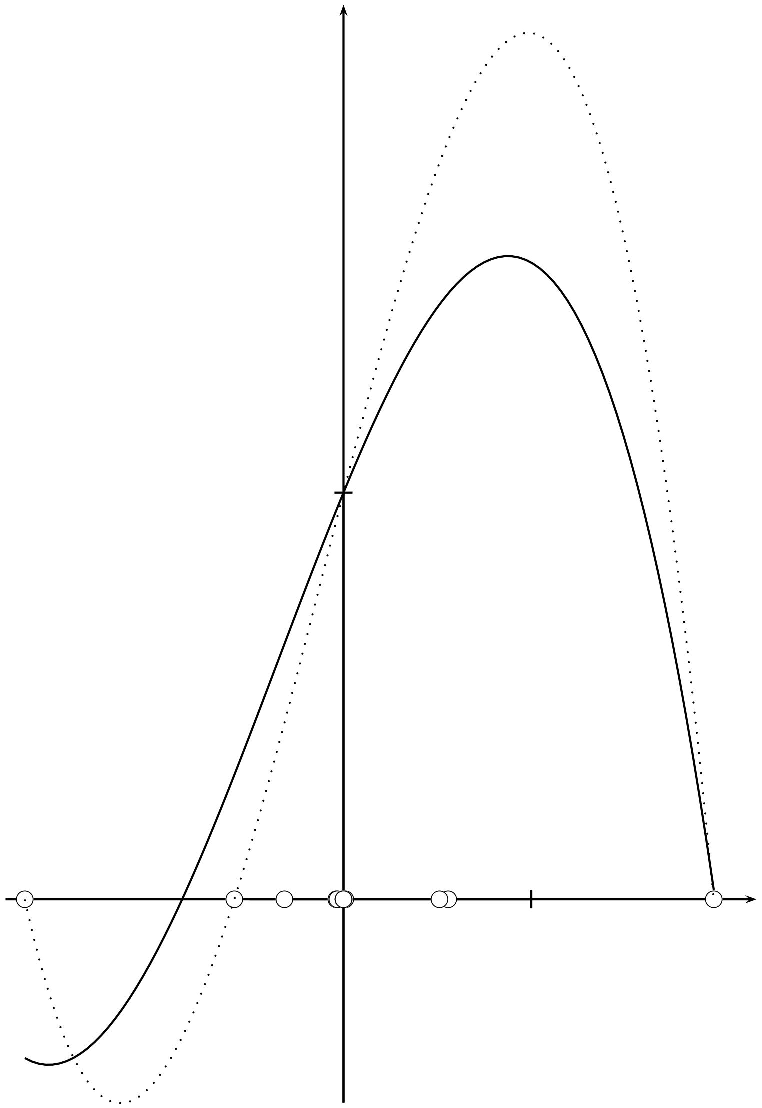
So the singular values and eigenvalues are just the reciprocal of before. This means that we also have the inequality

$$\prod_{n=1}^{N} \frac{1}{|\lambda_n|} \leq \prod_{n=1}^{N} \frac{1}{\mu_n} \quad \Leftrightarrow \quad \prod_{n=1}^{N} \mu_n \leq \prod_{n=1}^{N} |\lambda_n|. \qquad (6.18)$$

So, combined with the inequality in (6.17), we have for finite dimensional, invertible operators:

$$\prod_{n=1}^{N} |\lambda_n| = \prod_{n=1}^{N} \mu_n.$$

This equality also holds when K is not invertible. In that case, each side of the above expression is just zero.



CHAPTER 7

Krylov Subspace Methods

method, *the procedures and techniques characteristic of a particular discipline or field of knowledge*
— THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE

We now turn the focus to Krylov subspace methods. The cornerstone of these is not surprisingly Krylov subspaces.

Definition 7.1 *The* ***Krylov subspace*** $\mathcal{K}_k(A, r)$ *is defined as*

$$\mathcal{K}_k(A, r) = \text{span}\{r, Ar, \ldots, A^{k-1}r\}.$$

Given an equation, $Ax = b$, the goal of these methods is now to assemble a best solution, in some sense, from $\mathcal{K}_k(A, r)$ for each iteration step k. The vector r is typically the right-hand side $r = b$.

But how good a subspace is it to get solutions from? This is the main subject of the first part of this chapter.

7.1 Operator Smoothing

As discussed in Section 5.2, an integral operator tends to smoothen the input. But will the image always be continuous? No. Consider the kernel $k \in L^2([0,1]^2)$ defined as

$$k(s,t) = \begin{cases} 1, & \text{for } 0 \le s \le \frac{1}{2} \text{ and } 0 \le t \le 1 \\ 0, & \text{otherwise.} \end{cases}$$

Then we have for $f = 1$:

$$\int_0^1 k(s,t)f(t)dt = 1_{[0,\frac{1}{2}]},$$

which is obviously not continuous.

Let us consider an arbitrary kernel $k \in L^2(I \times J)$ and input $f \in L^2(J)$. For $g(s) = Kf(s)$ to be continuous, we must have: For all $s_0 \in I$ and all $\epsilon > 0$ there must exist a $\delta > 0$ such that for all $s \in I$,

$$|s - s_0| < \delta \quad \Rightarrow \quad |g(s) - g(s_0)| < \epsilon,$$

which in this setting corresponds to

$$|s - s_0| < \delta \quad \Rightarrow \quad \left| \int_J (k(s,t) - k(s_0,t))f(t)dt \right| < \epsilon$$

If $k(s,\cdot) \to k(s_0,\cdot)$ when $s \to s_0$ for all $s_0 \in I$, the image g is guaranteed to be continuous because of the Cauchy-Schwartz inequality. Another way of saying this: If g is *not* continuous then there must exist an $s_0 \in I$ such that $k(s,\cdot)$ does not converge to $k(s_0,\cdot)$ when $s \to s_0$.

The discontinuity of Volterra kernels are mostly along the diagonal, and so the image will still be continuous. No integral kernel from "real life problems" known to the author contains such a discontinuity. This means that every Krylov subspace will be spanned only by continuous vectors, with the possible exception of the starting vector (r in Definition 7.1). In turn, this means that the solutions are bound to become smooth.

Note that this also means that eigenvectors will *always* be smooth (unless the kernel has that special discontinuity).

7.2 The Existence of Solutions

Given an equation $Ax = b$, does the solution x lie in $\mathcal{K}_k(A,b)$ for some k? What if A is not invertible? In the following we will answer these questions for $\mathbf{Ax} = \mathbf{b}$, the finite dimensional (matrix) case.[1]

A useful tool here is the *Jordan decomposition*. If $\mathbf{A} \in \mathbb{C}^{n\times n}$ then there exists a factorization $\mathbf{X}^{-1}\mathbf{AX} = \operatorname{diag}(\mathbf{J}^{(1)}, \mathbf{J}^{(2)}, \dots, \mathbf{J}^{(r)})$ where $\mathbf{X} \in \mathbb{C}^{n\times n}$ is nonsingular and

$$\mathbf{J}^{(i)} = \begin{bmatrix} \lambda_i & 1 & & \cdots & 0 \\ 0 & \lambda_i & \ddots & & \vdots \\ & \ddots & \ddots & \ddots & \\ \vdots & & \ddots & \ddots & 1 \\ 0 & \cdots & & 0 & \lambda_i \end{bmatrix} \in \mathbb{C}^{n_i \times n_i}$$

where $n_1 + n_2 + \cdots + n_r = n$. Note that this decomposition is well-defined for *all* matrices and that it is a generalization of diagonalization.[2]

7.2.1 The Minimal Polynomial

As we shall see, Krylov subspace methods are intimately tied to the *minimal polynomial*. The minimal polynomial q of $\mathbf{A}$ is defined as the unique monic polynomial of minimal degree such that $q(\mathbf{A}) = \mathbf{0}$.

Given the Jordan decomposition, as defined above, we actually have an explicit expression for the minimal polynomial. Let the distinct eigenvalues of $\mathbf{A}$ be $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_d$ and let $\hat{\lambda}_i$ have *index* m_i. The index of an eigenvalue $\hat{\lambda}_i$ is defined as the largest Jordan block associ-

[1] See [IM98] for related results.

[2] See e.g. [GvL96] for additional information on Jordan decomposition.

ated with $\hat{\lambda}_i$. The minimal polynomial can now be expressed as[3]

$$q(\lambda) = \prod_{j=1}^{d} (\lambda - \hat{\lambda}_j)^{m_j}. \tag{7.1}$$

We also define m as the degree of the minimal polynomial, $m = \sum_{j=1}^{d} m_j$. Since $\mathbf{A}$ can have at most n distinct eigenvalues we have $m \leq n$. Consider for instance the matrix

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & & & & & \\ & 2 & & & & & \\ & & 3 & & & & \\ & & & 3 & & & \\ & & & & 2 & 1 & \\ & & & & & 2 & 1 \\ & & & & & & 2 \end{bmatrix},$$

which is already on Jordan form, consisting of 4 blocks. It has an eigenvalue 2 with index 3 and an eigenvalue 3 with index 1. The minimal polynomial in this case is $q(\lambda) = (\lambda - 2)^3(\lambda - 3)$.

In general when $\mathbf{A}$ is diagonalizable, m is equal to the number of *distinct* eigenvalues.

From Equation (7.1) we get the alternative expression

$$q(\lambda) = \sum_{j=0}^{m} \alpha_j \lambda^j,$$

where $\alpha_m = 1$ (it is monic) and $\alpha_0 = \prod_{j=1}^{d} (-\hat{\lambda}_j)^{k_j}$. So $\mathbf{A}$ is nonsingular if and only if $\alpha_0 \neq 0$.

A direct consequence of the definition of the minimal polynomial is that $q(\mathbf{A})\mathbf{b} = \sum_{j=0}^{m} \alpha_j \mathbf{A}^j \mathbf{b} = \mathbf{0}$ for all $\mathbf{b}$. This clearly shows

Theorem 7.2 *The vectors* $\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^m\mathbf{b}$, *where* m *is the degree of the minimal polynomial, are* linearly dependent *for all* $\mathbf{b}$.

This fact will come in handy later on.

[3] That the shown polynomial $q(\lambda)$ is monic and fulfills $q(\mathbf{A}) = \mathbf{0}$ is straightforward. See Section 2.8 in [Nev93] for a proof that q is actually *minimal*.

7.2.2 The Nonsingular Case

Assume that $\mathbf{A}$ is nonsingular, so $\alpha_0 \neq 0$. We now get

$$\begin{aligned} 0 = q(\mathbf{A}) = \alpha_0 \mathbf{I} + \alpha_1 \mathbf{A} + \cdots + \alpha_m \mathbf{A}^m \quad &\Leftrightarrow \\ \mathbf{I} = -\frac{1}{\alpha_0}\left(\alpha_1 \mathbf{I} + \cdots + \alpha_m \mathbf{A}^{m-1}\right)\mathbf{A} \quad &\Leftrightarrow \\ \mathbf{A}^{-1} = -\frac{1}{\alpha_0} \sum_{j=0}^{m-1} \alpha_{j+1} \mathbf{A}^j. & \end{aligned} \tag{7.2}$$

So the inverse of every nonsingular matrix $\mathbf{A}$ can be written as a polynomial in $\mathbf{A}$ itself. Furthermore,

$$\mathbf{x} = \mathbf{A}^{-1}\mathbf{b} = -\frac{1}{\alpha_0} \sum_{j=0}^{m-1} \alpha_{j+1} \mathbf{A}^j \mathbf{b} \in \mathcal{K}_m(\mathbf{A}, \mathbf{b}),$$

or written in words: For every equation $\mathbf{A}\mathbf{x} = \mathbf{b}$ with nonsingular $\mathbf{A}$, the solution $\mathbf{x}$ lies in $\mathcal{K}_m(\mathbf{A}, \mathbf{b})$.

7.2.3 The Singular Case

Let us now consider the singular case. Assume the Jordan decomposition is given as $\mathbf{A} = \mathbf{X}\mathbf{J}\mathbf{X}^{-1}$. We now transform into the basis given by $\mathbf{X}$,

$$\mathbf{A}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{X}\mathbf{J}\mathbf{X}^{-1}\mathbf{x} = \mathbf{b} \quad \Leftrightarrow \quad \mathbf{J}\boldsymbol{\xi} = \boldsymbol{\beta}, \tag{7.3}$$

where $\boldsymbol{\xi} = \mathbf{X}^{-1}\mathbf{x}$ and $\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{b}$.

Since $\mathbf{A}$ is singular we can split $\mathbf{J}$ so that

$$\mathbf{J} = \begin{bmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0} & \mathbf{N} \end{bmatrix},$$

where all the zero eigenvalues are contained in $\mathbf{N}$ (is always possible by appropriately exchanging columns in $\mathbf{X}$). From this follows that $\mathbf{C}$ is regular and that $\mathbf{N}$ is *nilpotent*. Nilpotent means that there exists an integer i, called the index, such that $\mathbf{N}^i = \mathbf{0}$ while $\mathbf{N}^{i-1} \neq \mathbf{0}$.

Fortunately, the nilpotent index coincides with the index of the zero eigenvalue.

Suppose that a Krylov solution exists, that is for some p we have

$$\mathbf{x} = \sum_{j=0}^{p} \alpha_j \mathbf{A}^j \mathbf{b} \quad \Leftrightarrow \quad \boldsymbol{\xi} = \sum_{j=0}^{p} \alpha_j \mathbf{J}^j \boldsymbol{\beta} = \sum_{j=0}^{p} \alpha_j \begin{bmatrix} \mathbf{C}^j & \mathbf{0} \\ \mathbf{0} & \mathbf{N}^j \end{bmatrix} \boldsymbol{\beta}. \tag{7.4}$$

If we now let $\boldsymbol{\xi} = \begin{bmatrix} \boldsymbol{\xi}_C \\ \boldsymbol{\xi}_N \end{bmatrix}$ and $\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_C \\ \boldsymbol{\beta}_N \end{bmatrix}$, where the splitting follows that of $\mathbf{J}$ we get

$$\boldsymbol{\xi}_C = \sum_{j=0}^{p} \alpha_j \mathbf{C}^j \boldsymbol{\beta}_C \quad \text{and} \quad \boldsymbol{\xi}_N = \sum_{j=0}^{p} \alpha_j \mathbf{N}^j \boldsymbol{\beta}_N.$$

Since $\mathbf{J}\boldsymbol{\xi} = \boldsymbol{\beta}$ we also have $\mathbf{N}\boldsymbol{\xi}_N = \boldsymbol{\beta}_N$ so

$$\mathbf{N}\left(\sum_{j=0}^{p} \alpha_j \mathbf{N}^j \boldsymbol{\beta}_N\right) = \boldsymbol{\beta}_N \quad \Leftrightarrow \quad \left(\mathbf{I} - \sum_{j=0}^{p} \alpha_j \mathbf{N}^{j+1}\right) \boldsymbol{\beta}_N = \mathbf{0}.$$

The matrices $\mathbf{N}^j$, $j \geq 2$, are upper triangular and have only zeros along the diagonal. This implies that the matrix in the right-most parenthesis above will be regular. So if a Krylov solution exists we must have $\boldsymbol{\beta}_N = \mathbf{0}$. Another way of saying this is that we must have

$$\boldsymbol{\beta} \in \mathcal{R}\left(\begin{bmatrix} \mathbf{C} & \\ & \mathbf{0} \end{bmatrix}\right) = \mathcal{R}\left(\begin{bmatrix} \mathbf{C}^i & \\ & \mathbf{N}^i \end{bmatrix}\right) = \mathcal{R}\left(\mathbf{J}^i\right),$$

where $\mathcal{R}(\mathbf{C}) = \mathcal{R}(\mathbf{C}^i)$ is valid because $\mathbf{C}$ is regular.

Let us consider the converse case and assume $\boldsymbol{\beta} \in \mathcal{R}(\mathbf{J}^i)$, that is,

$$\boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_C \\ \mathbf{0} \end{bmatrix} \quad \text{which implies} \quad \boldsymbol{\xi} = \begin{bmatrix} \mathbf{C}^{-1}\boldsymbol{\beta}_C \\ \mathbf{0} \end{bmatrix}.$$

The matrix $\mathbf{C}$ is regular and its minimal polynomial has degree $m-i$. This means that there exists a polynomial $q(\mathbf{C})$ of degree $m-i-1$ such that $q(\mathbf{C}) = \mathbf{C}^{-1}$. Then we get

$$\begin{aligned} \boldsymbol{\xi} &= \begin{bmatrix} \mathbf{C}^{-1}\boldsymbol{\beta}_C \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} q(\mathbf{C}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_C \\ \mathbf{0} \end{bmatrix} = \begin{bmatrix} q(\mathbf{C}) & \mathbf{0} \\ \mathbf{0} & q(\mathbf{N}) \end{bmatrix} \boldsymbol{\beta} \\ &= q(\mathbf{J})\boldsymbol{\beta} \in \mathcal{K}_{m-i}(\mathbf{J}, \boldsymbol{\beta}). \end{aligned}$$

What we have now is that a Krylov solution exists for the system $\mathbf{J}\boldsymbol{\xi} = \boldsymbol{\beta}$ if and only if $\boldsymbol{\beta} \in \mathcal{R}(\mathbf{J}^i)$, where i is the index of the zero eigenvalue. But the existence of a Krylov solution to $\mathbf{J}\boldsymbol{\xi} = \boldsymbol{\beta}$ and $\mathbf{A}\mathbf{x} = \mathbf{b}$ is equivalent, cf. (7.4), and furthermore $\mathcal{R}(\mathbf{J}^i) = \mathcal{R}(\mathbf{A}^i)$ since $\mathbf{X}$ is regular.

7.2.4 In Summation

The above statements lead to this important theorem:

Theorem 7.3 *A Krylov solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ exists if and only if $\mathbf{b} \in \mathcal{R}(\mathbf{A}^i)$, where i is the index of the zero eigenvalue of $\mathbf{A}$.*

Notice that this statement is equally valid when $\mathbf{A}$ is nonsingular since we here have $i = 0$.

We have furthermore shown, for general $\mathbf{A}$, that

Theorem 7.4 *When a Krylov solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ exists we have*

$$\mathbf{x} \in \mathcal{K}_{m-i}(\mathbf{A}, \mathbf{b}),$$

where m is the degree of the minimal polynomial and i is the index of the zero eigenvalue.

When $\mathbf{A}$ is diagonalizable these theorems can be simplified. The degree of the minimal polynomial becomes the number of distinct eigenvalues, and $i = 1$ if $\mathbf{A}$ is singular and $i = 0$ otherwise. From this follows for *diagonalizable* $\mathbf{A}$:

- A Krylov solution exists if and only if $\mathbf{b} \in \mathcal{R}(\mathbf{A})$.
- When a Krylov solution exists we have $\mathbf{x} \in \mathcal{K}_{d_0}(\mathbf{A}, \mathbf{b})$ where d_0 is the number of distinct eigenvalues *different from* 0.

7.3 Detection of a Solution

If some method iteratively creates Krylov subspaces of larger and larger dimension, how does one know when to stop? When is the subspace large enough to assemble a solution, and can it be detected that no Krylov solution exists at all?

Before answering these question, we prove an elementary, but important, fact about Krylov subspaces.

Theorem 7.5 *If $\mathbf{A}^k\mathbf{b} \in \mathcal{K}_k(\mathbf{A},\mathbf{b})$, then $\mathbf{A}^j\mathbf{b} \in \mathcal{K}_k(\mathbf{A},\mathbf{b})$ for all $j \geq k$.*

Proof: From the assumptions we have

$$\mathbf{A}^k\mathbf{b} = \alpha_1\mathbf{b} + \alpha_2\mathbf{A}\mathbf{b} + \cdots + \alpha_k\mathbf{A}^{k-1}\mathbf{b} \in \mathcal{K}_k(\mathbf{A},\mathbf{b}).$$

Then we have

$$\begin{aligned}\mathbf{A}^{k+1}\mathbf{b} &= \mathbf{A}(\mathbf{A}^k\mathbf{b}) = \mathbf{A}(\alpha_1\mathbf{b} + \alpha_2\mathbf{A}\mathbf{b} + \cdots + \alpha_k\mathbf{A}^{k-1}\mathbf{b}) \\ &= \alpha_1\mathbf{A}\mathbf{b} + \alpha_2\mathbf{A}^2\mathbf{b} + \cdots + \alpha_k\mathbf{A}^k\mathbf{b} \in \mathcal{K}_k(\mathbf{A},\mathbf{b}).\end{aligned}$$

From induction the result follows. □

If $\dim(\mathcal{K}_k(\mathbf{A},\mathbf{b})) = k$ and $\dim(\mathcal{K}_{k+1}(\mathbf{A},\mathbf{b})) < k+1$ then it follows from this theorem that we must have $\dim(\mathcal{K}_j(\mathbf{A},\mathbf{b})) = k$ for all $j > k$.

Assume now that a Krylov solution has been found, $\mathbf{x} \in \mathcal{K}_k(\mathbf{A},\mathbf{b})$, where the Krylov subspace has dimension k. Written explicitly,

$$\mathbf{x} = \alpha_1\mathbf{b} + \alpha_2\mathbf{A}\mathbf{b} + \cdots + \alpha_k\mathbf{A}^{k-1}\mathbf{b},$$

where $\alpha_k \neq 0$. Then we have

$$\mathbf{0} = \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{b} - \alpha_1\mathbf{A}\mathbf{b} - \alpha_2\mathbf{A}^2\mathbf{b} - \cdots - \alpha_k\mathbf{A}^k\mathbf{b}.$$

So the vectors $\mathbf{b}, \mathbf{A}\mathbf{b}$, ..., $\mathbf{A}^k\mathbf{b}$ are linearly dependent, or equivalently, $\dim(\mathcal{K}_{k+1}(\mathbf{A},\mathbf{b})) = k$. This implies that the *only* linear combination of the vectors $\mathbf{b}$, $\mathbf{A}\mathbf{b}$, ..., $\mathbf{A}^k\mathbf{b}$ that gives the zero vector is the combination shown above (when fixing the coefficient to $\mathbf{b}$).

We have now proved the following theorem.

Theorem 7.6 *If a Krylov solution $\mathbf{x} \in \mathcal{K}_k(\mathbf{A},\mathbf{b})$ exists where the Krylov subspace has dimension k, then $\dim(\mathcal{K}_{k+1}(\mathbf{A},\mathbf{b})) = k$. Furthermore, if*

$$\alpha_0\mathbf{b} + \alpha_1\mathbf{A}\mathbf{b} + \alpha_2\mathbf{A}^2\mathbf{b} + \cdots + \alpha_k\mathbf{A}^k\mathbf{b} = \mathbf{0}$$

where not all α_i are zero, we have that both α_0 and α_k are non-zero.

This theorem states exactly when we should stop to look for a solution. If $\dim(\mathcal{K}_k(\mathbf{A},\mathbf{b})) = k$ and $\dim(\mathcal{K}_{k+1}(\mathbf{A},\mathbf{b})) < k+1$ the solution, *if it exists*, lies in $\mathcal{K}_k(\mathbf{A},\mathbf{b})$.

But what about the existence? The following theorem has a result on that.

Theorem 7.7 *Assume that a non-zero vector $\mathbf{w} \in \mathcal{K}_k(\mathbf{A},\mathbf{b})$ exists such that $\mathbf{A}\mathbf{w} = \mathbf{0}$ and $\dim(\mathcal{K}_k(\mathbf{A},\mathbf{b})) = k$. Then $\mathbf{A}$ is singular and no Krylov solution exists.*

Proof: That $\mathbf{A}$ is singular is straightforward. Let

$$\mathbf{w} = \beta_1\mathbf{b} + \beta_2\mathbf{A}\mathbf{b} + \cdots + \beta_k\mathbf{A}^{k-1}\mathbf{b},$$

where $\beta_k \neq 0$, which implies

$$\mathbf{A}\mathbf{w} = \beta_1\mathbf{A}\mathbf{b} + \beta_2\mathbf{A}^2\mathbf{b} + \cdots + \beta_k\mathbf{A}^k\mathbf{b} = \mathbf{0}. \tag{7.5}$$

Assume that a Krylov solution in $\mathcal{K}_j(\mathbf{A},\mathbf{b})$ exists where the dimension of $\mathcal{K}_j(\mathbf{A},\mathbf{b})$ is j. According to Theorem 7.6 we have

$$\mathbf{b} + \alpha_1\mathbf{A}\mathbf{b} + \cdots + \alpha_j\mathbf{A}^j\mathbf{b} = \mathbf{0}, \tag{7.6}$$

for some constants α_i, $i = 1, 2, \ldots, j$, where $\alpha_j \neq 0$. Because of this linear dependency and since $\dim(\mathcal{K}_k(\mathbf{A},\mathbf{b})) = k$ we must have $j \geq k$. Because of Equation (7.5) and $\dim(\mathcal{K}_j(\mathbf{A},\mathbf{b})) = j$ we also have $j \leq k$. Hence $j = k$. If we now subtract α_k/β_k times (7.5) from (7.6) we see that we have a non-trivial linear dependency among the vectors $\mathbf{b}, \mathbf{A}\mathbf{b}, \ldots, \mathbf{A}^{k-1}\mathbf{b}$ which is a contradiction. So the assumption that a Krylov solution existed was false. □

The usability of this theorem may not be obvious, but we shall see in the next chapter how an implementation of GMRES can use this result.

7.4 GMRES

Let us now look at GMRES using the notation introduced for iterative methods in Section 5.4. At each iteration k, the approximate solution $\mathbf{x}^{(k)}$ found by GMRES fulfills:

$$\mathbf{x}^{(k)} = \operatorname*{argmin}_{\mathbf{x}\in\mathbf{x}^{(0)}+\mathcal{K}_k(\mathbf{A},\mathbf{r}^{(0)})} \|\mathbf{r}(\mathbf{x})\|_2, \tag{7.7}$$

where $\mathbf{x}^{(0)}$ is some starting vector (this means that we actually have to solve $\mathbf{A}\mathbf{z} = \mathbf{r}^{(0)}$ instead and then compute $\mathbf{x} = \mathbf{z} + \mathbf{x}^{(0)}$).

Since $\mathcal{K}_1(\mathbf{A},\mathbf{r}^{(0)}) \subset \mathcal{K}_2(\mathbf{A},\mathbf{r}^{(0)}) \subset \cdots$, it is clear that the residual norm can only decrease from one iteration to the next.

7.4.1 Convergence Analysis

The notation $\mathbf{x} \in \mathbf{x}^{(0)} + \mathcal{K}_k(\mathbf{A},\mathbf{r}^{(0)})$ should be understood as $\mathbf{x} = \mathbf{x}^{(0)} + \mathbf{z}$ where $\mathbf{z} \in \mathcal{K}_k(\mathbf{A},\mathbf{r}^{(0)})$. So each possible $\mathbf{x}$ can be expressed as

$$\mathbf{x} = \mathbf{x}^{(0)} + \sum_{i=0}^{k-1} \mathbf{c}_{i+1}^{(k)}\mathbf{A}^i\mathbf{r}^{(0)},$$

for some vector $\mathbf{c}^{(k)} \in \mathbb{C}^k$. The residual vector becomes:

$$\begin{aligned}\mathbf{r}(\mathbf{x}) &= \mathbf{b} - \mathbf{A}\mathbf{x} = \mathbf{b} - \mathbf{A}\mathbf{x}^{(0)} - \mathbf{A}\sum_{i=0}^{k-1} \mathbf{c}_{i+1}^{(k)}\mathbf{A}^i\mathbf{r}^{(0)} \\ &= \mathbf{r}^{(0)} - \sum_{i=1}^{k} \mathbf{c}_i^{(k)}\mathbf{A}^i\mathbf{r}^{(0)} = \left(\mathbf{I} - \sum_{i=1}^{k} \mathbf{c}_i^{(k)}\mathbf{A}^i\right)\mathbf{r}^{(0)} = p_k(\mathbf{A})\mathbf{r}^{(0)},\end{aligned} \tag{7.8}$$

where p_k is a polynomial of maximal degree k and for which $p_k(0) = 1$. Let $\mathcal{P}_k$ denote the set of all such polynomials. Note that p_k and $\mathbf{c}^{(k)}$ are implicitly related by

$$\begin{aligned}p_k(\lambda) &= 1 - \mathbf{c}_1^{(k)}\lambda - \mathbf{c}_2^{(k)}\lambda^2 - \cdots - \mathbf{c}_k^{(k)}\lambda^k \\ &= 1 - \begin{bmatrix}\lambda & \lambda^2 & \cdots & \lambda^k\end{bmatrix}\mathbf{c}^{(k)}.\end{aligned} \tag{7.9}$$

So for every $\mathbf{x} \in \mathbf{x}^{(0)} + \mathcal{K}_k(\mathbf{A},\mathbf{r}^{(0)})$ we have $\mathbf{r}(\mathbf{x}) = p_k(\mathbf{A})\mathbf{r}^{(0)}$ for some $p_k \in \mathcal{P}_k$. This means that

$$\|\mathbf{r}^{(k)}\|_2 = \|\mathbf{r}(\mathbf{x}^{(k)})\|_2 \le \|p_k(\mathbf{A})\mathbf{r}^{(0)}\|_2 \quad \text{for all } p_k \in \mathcal{P}_k, \tag{7.10}$$

since GMRES in each iteration finds the *optimal* solution expressed by (7.7). The polynomial corresponding to the optimal solution will be denoted p_k^{opt},

$$\|\mathbf{r}^{(k)}\|_2 = \|p_k^{\text{opt}}(\mathbf{A})\mathbf{r}^{(0)}\|_2 = \min_{p_k\in\mathcal{P}_k} \|p_k(\mathbf{A})\mathbf{r}^{(0)}\|_2.$$

A typical way to provide an upper bound for $\|\mathbf{r}^{(k)}\|_2$ is to use the inequality $\|p_k(\mathbf{A})\mathbf{r}^{(0)}\|_2 \le \|p_k(\mathbf{A})\|_2\|\mathbf{r}^{(0)}\|_2$ and then investigate how small $\|p_k(\mathbf{A})\|_2$ can be. This has lead to some very good convergence bounds for GMRES, but it clearly neglects all information about the right-hand side. This is unfortunate since we have seen earlier that solving discrete ill-posed problems can be extremely sensitive to the right-hand side. So we will try to look at the whole expression, $\|p_k(\mathbf{A})\mathbf{r}^{(0)}\|_2$, instead.

Using the Spectral Decomposition

In order to be able to proceed we will make the assumption that $\mathbf{A}$ is diagonalizable, that is, there exists a regular matrix $\mathbf{X} \in \mathbb{C}^{n\times n}$ such that

$$\mathbf{A} = \mathbf{X}\boldsymbol{\Lambda}\mathbf{X}^{-1}, \quad \boldsymbol{\Lambda} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n),$$

where $|\lambda_1| \ge |\lambda_2| \ge \cdots \ge |\lambda_n|$. Note that no loss of generality is suffered by requiring the eigenvalues must be sorted in the given order.

Since $\mathbf{A}^i = \mathbf{X}\boldsymbol{\Lambda}^i\mathbf{X}^{-1}$ we get from the decomposition,

$$\mathbf{r}(\mathbf{x}) = p_k(\mathbf{A})\mathbf{r}^{(0)} = \mathbf{X}p_k(\boldsymbol{\Lambda})\mathbf{X}^{-1}\mathbf{r}^{(0)}.$$

We now introduce $\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{r}^{(0)}$, $\operatorname{diag}(\mathbf{L}_i) = \boldsymbol{\Lambda}^i$ and $\operatorname{diag}(\mathbf{e}) = \mathbf{I}$. If we furthermore use the obvious equality,

$$\operatorname{diag}(\mathbf{v})\mathbf{w} = \begin{bmatrix}\mathbf{v}_1\mathbf{w}_1 & \mathbf{v}_1\mathbf{w}_1 & \cdots & \mathbf{v}_n\mathbf{w}_n\end{bmatrix}^T = \operatorname{diag}(\mathbf{w})\mathbf{v},$$

we rewrite as follows

$$\begin{aligned}
\mathbf{X} p_k(\mathbf{\Lambda})\mathbf{X}^{-1}\mathbf{r}^{(0)} &= \mathbf{X}\left(\mathbf{I} - \sum_{i=1}^{k} \mathbf{c}_i^{(k)} \mathbf{\Lambda}^i\right)\beta \\
&= \mathbf{X}\mathrm{diag}\left(\mathbf{e} - \sum_{i=1}^{k} \mathbf{c}_i^{(k)} \mathbf{L}_i\right)\boldsymbol{\beta} = \mathbf{X}\mathrm{diag}(\boldsymbol{\beta})\left(\mathbf{e} - \mathbf{L}\mathbf{c}^{(k)}\right). \qquad (7.11)
\end{aligned}$$

So the bound in Equation (7.10) can be expressed as

$$\begin{aligned}
\|\mathbf{r}^{(k)}\|_2 &\le \|\mathbf{X}\|_2 \left\|\mathrm{diag}(\boldsymbol{\beta})\left(\mathbf{e} - \mathbf{L}\mathbf{c}^{(k)}\right)\right\|_2 \\
&= \|\mathbf{X}\|_2 \left\| \begin{bmatrix} \boldsymbol{\beta}_1 & & & \\ & \boldsymbol{\beta}_2 & & \\ & & \ddots & \\ & & & \boldsymbol{\beta}_n \end{bmatrix} \left(\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^k \\ \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^k \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_n & \lambda_n^2 & \cdots & \lambda_n^k \end{bmatrix} \mathbf{c}^{(k)} \right) \right\|_2 \qquad (7.12)
\end{aligned}$$

for all $\mathbf{c}^{(k)} \in \mathbb{C}^k$.

The matrix $\mathbf{L}$ is seen to be similar to a transposed Vandermonde matrix, often used for polynomial interpolation.[4]

Minimizing the norm in Equation (7.12) is frequently called a *weighted least squares problem*. If $\boldsymbol{\beta}_i = 1$ for all i it is seen to be a usual least squares problem, but weighting the ith row/equation with a large $\boldsymbol{\beta}_i$ makes it more important to fulfill this particular equation.

[4]Given elements $v_0, v_1, \ldots, v_n$, a Vandermonde matrix $\mathbf{V}$ has entries $\mathbf{V}_{i+1,j+1} = v_j^i$ for $i, j = 0, 1, \ldots, n$. See more in e.g. [GvL96] page 183.

Interpretation in Terms of Polynomials

Inequality (7.12) can also be written using the relation (7.9), obtaining the equivalent

$$\begin{aligned}
\|\mathbf{r}^{(k)}\|_2^2 \le& \|\mathbf{X}\|_2^2 \sum_{i=1}^{n} |\boldsymbol{\beta}_i|^2 \left|1 - \left[\lambda_i \ \lambda_i^2 \ \cdots \ \lambda_i^k\right] \mathbf{c}^{(k)}\right|^2 \\
&= \|\mathbf{X}\|_2^2 \sum_{i=1}^{n} |\boldsymbol{\beta}_i|^2 |p_k(\lambda_i)|^2
\end{aligned} \qquad (7.13)$$

for all $p_k \in \mathcal{P}_k$. Note that this bound depends only on the *absolute value* of each element $\boldsymbol{\beta}_i$ and the magnitude of p_k in each eigenvalue λ_i. We introduce the function

$$w(\lambda) = \sum_{\lambda_i = \lambda} |\boldsymbol{\beta}_i|^2.$$

The inequality in (7.13) can now be written as

$$\begin{aligned}
\|\mathbf{r}^{(k)}\|_2^2 \le& \|\mathbf{X}\|_2^2 \sum_{i=1}^{d} w(\hat{\lambda}_i)\left|p_k(\hat{\lambda}_i)\right|^2 \\
&= \|\mathbf{X}\|_2^2 \left(w(0) + \sum_{i=1}^{d_0} w(\hat{\lambda}_i)\left|p_k(\hat{\lambda}_i)\right|^2\right)
\end{aligned} \qquad (7.14)$$

where $|\hat{\lambda}_1| > |\hat{\lambda}_2| > \cdots > |\hat{\lambda}_{d_0}| > 0$ are the distinct eigenvalues different from zero.

This expression indicates that when the right-hand side has components in the null-space of $\mathbf{A}$ we get $w(0) \neq 0$ and the residual can never vanish.

We now wish to bound the term

$$\sum_{i=1}^{d_0} w(\hat{\lambda}_i)\left|p_k(\hat{\lambda}_i)\right|^2,$$

remembering that every choice of $p_k \in \mathcal{P}_k$ will provide us with a valid bound. We get the inspiration from Figure 7.1. Here, a model problem with real eigenvalues has been created and iteration $1, 2, 3, 4$ of GMRES is represented. In each plot, the optimal polynomial p_k^{opt} is shown together with the polynomial $p_k^{\max}$ that interpolates the k largest (in absolute value) eigenvalues. This can be written explicitly as

$$p_k^{\max}(\lambda) = \prod_{i=1}^{k} \left(1 - \frac{\lambda}{\hat{\lambda}_i}\right), \tag{7.15}$$

for $k = 1, 2, \ldots, d_0$. By inserting this expression into the bound in Equation (7.14) we have a valid, and hopefully useful, upper bound on the residual.

Approximating the Convergence Rate

We will now make the assumption that the eigenvalues of $\mathbf{A}$ are all non-zero and distinct. Then the residual bound becomes

$$\|\mathbf{r}^{(k)}\|_2^2 \leq \|\mathbf{X}\|_2^2 \sum_{i=k+1}^{n} |\boldsymbol{\beta}_i|^2 \prod_{j=1}^{k} \left|1 - \frac{\lambda_i}{\lambda_j}\right|^2 .$$

To make any convergence approximations using this expression, we must know both the eigenvalues λ_i and $\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{r}^{(0)}$. To simplify things, we use the approximation $p_k^{\max}(\lambda) \simeq 1$ for all λ. This can be justified because we have $p_k^{\max}(\lambda) \to 1$ as $\lambda \to 0$ and we only need values of $p_k^{\max}(\lambda)$ for very small λ. Now the bound has become

$$\|\mathbf{r}^{(k)}\|_2^2 \lesssim \|\mathbf{X}\|_2^2 \sum_{i=k+1}^{n} |\boldsymbol{\beta}_i|^2 . \tag{7.16}$$

We now assume that the right-hand side has eigenvector components decreasing exponentially as $|\boldsymbol{\beta}_i| = d \cdot q^{-i}$, and consider the quantity

$$(s^{(k)})^2 = \sum_{i=k+1}^{n} |\boldsymbol{\beta}_i|^2 = d^2 \sum_{i=k+1}^{n} q^{-2i}.$$

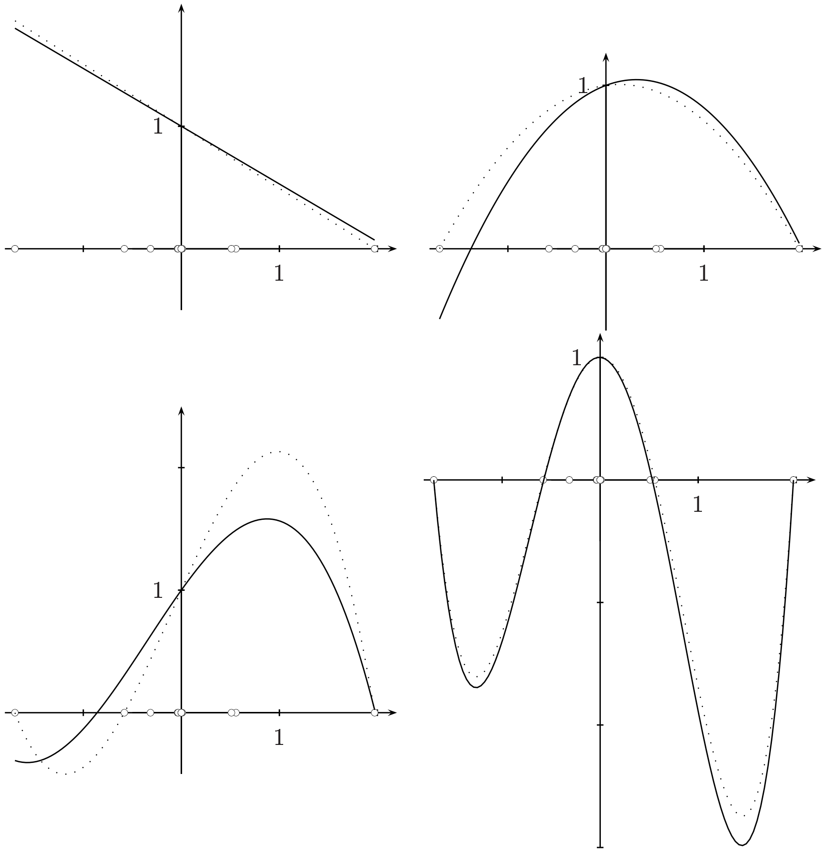


__Figure 7.1:__ The polynomial p_k^{opt} (solid line) corresponding to the optimal solution for iterations $k = 1, 2, 3, 4$ of GMRES. The dotted lines show the polynomials that interpolates the k largest (in absolute value) eigenvalues. The first coordinate of the dots represents the eigenvalues.

By calculating

$$(s^{(k)})^2(1-q^{-2}) = d^2\left(q^{-2(k+1)} - q^{-2(n+1)}\right) = d^2 q^{-2(k+1)}\left(1-q^{-2(n-k)}\right),$$

we see that

$$(s^{(k)})^2 = d^2 q^{-2(k+1)} \frac{1-q^{-2(n-k)}}{1-q^2}.$$

To see how this quantity changes from one iteration to the next, we calculate

$$\frac{(s^{(k+1)})^2}{(s^{(k)})^2} = \frac{q^{-2(k+2)}(1-q^{-2(n-k-1)})}{q^{-2(k+1)}(1-q^{-2(n-k)})} = q^{-2}\frac{1-q^{-2(n-k-1)}}{1-q^{-2(n-k)}} < q^{-2}.$$

Note that although this expression provides an upper bound, when $k \ll n$, we have "$\simeq$". If we now, compared to the approximate bound in Equation (7.16), assume that

$$\|\mathbf{r}^{(k)}\|_2 \simeq \|\mathbf{X}\|_2 s^{(k)},$$

we get an expression telling how the residual norm is reduced from one iteration to the next,

$$\frac{\|\mathbf{r}^{(k+1)}\|_2}{\|\mathbf{r}^{(k)}\|_2} \simeq q^{-1}, \quad \text{and hence} \quad \left(\frac{\|\mathbf{r}^{(k)}\|_2}{\|\mathbf{r}^{(0)}\|_2}\right)^{\frac{1}{k}} \simeq q^{-1},$$

for $k \ll n$.

Let us shortly return to the assumption concerning the asymptotic behavior of the $\boldsymbol{\beta}_i$'s. Is it realistic to know anything about them? Set $\mathbf{b} = \mathbf{A}\mathbf{x} + \mathbf{e}$ where $\mathbf{e}$ is some noise vector. We now get

$$\boldsymbol{\beta} = \mathbf{X}^{-1}\mathbf{b} = \mathbf{X}^{-1}\mathbf{A}\mathbf{x} + \mathbf{X}^{-1}\mathbf{e} = \boldsymbol{\Lambda}\mathbf{X}^{-1}\mathbf{x} + \mathbf{X}^{-1}\mathbf{e}$$

Consider the term $\boldsymbol{\Lambda}\mathbf{X}^{-1}\mathbf{x}$. If the eigenvector components of the solution, $\mathbf{X}^{-1}\mathbf{x}$, are of equal magnitude, the elements of $\boldsymbol{\Lambda}\mathbf{X}^{-1}\mathbf{x}$ will

decrease asymptotically like the eigenvalues. Assuming that the eigenvectors corresponding to large eigenvalues are slowly oscillating and that the solution is smooth, the elements of $\mathbf{X}^{-1}\mathbf{x}$ will decrease and the elements of $\boldsymbol{\Lambda}\mathbf{X}^{-1}\mathbf{x}$ will decrease even faster than the eigenvalues.

The term $\mathbf{X}^{-1}\mathbf{e}$ represents the influence of the noise. Assuming that the noise in $\mathbf{e}$ is white, i.e. contains (in principle) all wave frequencies with the same amplitude, the elements of $\mathbf{X}^{-1}\mathbf{e}$ will be approximately at the same level.

This means that the elements of $\boldsymbol{\beta}$ typically will decay at least as quickly as the eigenvalues, until they hit a certain error level.[5]

7.5 MINRES and CG

The Krylov subspace method MINRES is *identical* to GMRES in its mathematical formulation. The only difference lies in the implementation of the method, since it is an optimized version of GMRES that only works for symmetric coefficient matrices. So naturally, the convergence results above also apply to this method.

The Conjugate Gradients method, CG, is also a Krylov subspace method but for symmetric and positive definite[6] (SPD) coefficient matrices. It has a slightly different mathematical formulation than GMRES and MINRES. In [Gre97] it is proven for CG that the $\mathbf{A}$-norm of the error vector $\mathbf{e}^{(k)}$ is smallest among all vectors in the space

$$\mathbf{e}^{(0)} + \mathrm{span}\{\mathbf{A}\mathbf{e}^{(0)}, \mathbf{A}^2\mathbf{e}^{(0)}, \ldots, \mathbf{A}^k\mathbf{e}^{(0)}\}.$$

The $\mathbf{A}$-norm is defined as $\|\mathbf{v}\|_{\mathbf{A}} = (\mathbf{A}\mathbf{v}, \mathbf{v}) = \mathbf{v}^H\mathbf{A}\mathbf{v}$ and it can be proven well-defined since $\mathbf{A}$ is assumed to be SPD.

Let us denote the exact solution $\mathbf{x}^\star$ so we have $\mathbf{e}^{(k)} = \mathbf{x}^\star - \mathbf{x}^{(k)}$ and we remember that $\mathbf{A}\mathbf{e}^{(k)} = \mathbf{r}^{(k)}$. Then we get the following identical

[5] This resembles the behavior of the right-hand side Fourier coefficients shown in the Picard plot on page 65. Here, the coefficients were with respect to singular vectors, though.

[6] A matrix $\mathbf{A}$ is positive definite if $(\mathbf{A}\mathbf{v}, \mathbf{v}) = \mathbf{v}^H\mathbf{A}\mathbf{v} > 0$ for all non-null $\mathbf{v}$.

statement: In iteration k the $\mathbf{A}$-norm of the error $\mathbf{x}^\star - \mathbf{x}$ is minimal among all vectors $\mathbf{x}$ in the space

$$\mathbf{x}^{(0)} + \operatorname{span}\{\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \ldots, \mathbf{A}^{k-1}\mathbf{r}^{(0)}\}.$$

So the vector space that the solution is chosen from, is identical to that of GMRES (and MINRES). But how is the "best" solution in each iteration chosen? Since $\mathbf{A}$ is SPD it has no zero eigenvalues and is hence invertible. We then get from the minimization property of CG:

$$\|\mathbf{e}^{(k)}\|_{\mathbf{A}} = (\mathbf{A}\mathbf{e}^{(k)}, \mathbf{e}^{(k)}) = (\mathbf{e}^{(k)}, \mathbf{r}^{(k)}) = (\mathbf{A}^{-1}\mathbf{r}^{(k)}, \mathbf{r}^{(k)}) = \|\mathbf{r}^{(k)}\|_{\mathbf{A}^{-1}},$$

so it is equivalent to minimizing the $\mathbf{A}^{-1}$-norm of the residual instead! This brings us to a mathematical formulation of CG that is closely related to that of GMRES,

$$\mathbf{x}^{(k)} = \operatorname*{argmin}_{\mathbf{x}\in\mathbf{x}^{(0)}+\mathcal{K}_k(\mathbf{A},\mathbf{r}^{(0)})} \|\mathbf{r}(\mathbf{x})\|_{\mathbf{A}^{-1}}.$$

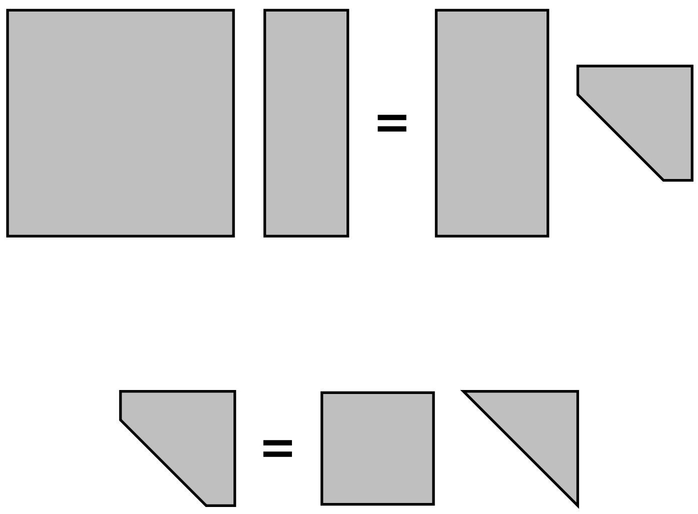
To proceed with the relation to the convergence analysis of GMRES, we first observe that $\mathbf{A}$ is symmetric so it has a decomposition of the form $\mathbf{A} = \mathbf{X}\mathbf{\Lambda}\mathbf{X}^T$ where $\mathbf{X}$ is orthogonal. Since it is also positive definite, it has positive eigenvalues and $\mathbf{A}^{-1} = \mathbf{X}\mathbf{\Lambda}^{-1}\mathbf{X}^T$ is well-defined. We also observe that for an arbitrary vector $\mathbf{v}$ we have

$$\|\mathbf{v}\|_{\mathbf{A}^{-1}} = \mathbf{v}^T\mathbf{X}\mathbf{\Lambda}^{-1}\mathbf{X}^T\mathbf{v} = \|\sqrt{\mathbf{\Lambda}^{-1}}\mathbf{X}^T\mathbf{v}\|_2.$$

Now we can write expressions analogous to those in (7.11), (7.12) and (7.13):

$$\begin{aligned}
&\|\mathbf{e}^{(k)}\|_{\mathbf{A}} = \|\mathbf{r}^{(k)}\|_{\mathbf{A}^{-1}} = \|\mathbf{X}p_k(\mathbf{\Lambda})\mathbf{X}^T\mathbf{r}^{(0)}\|_{\mathbf{A}^{-1}} = \|\sqrt{\mathbf{\Lambda}^{-1}}p_k(\mathbf{\Lambda})\boldsymbol{\beta}\|_2 \\
&= \left\| \begin{bmatrix} \boldsymbol{\beta}_1/\lambda_1 & & & \\ & \boldsymbol{\beta}_2/\lambda_2 & & \\ & & \ddots & \\ & & & \boldsymbol{\beta}_n/\lambda_n \end{bmatrix} \left(\begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} - \begin{bmatrix} \lambda_1 & \lambda_1^2 & \cdots & \lambda_1^k \\ \lambda_2 & \lambda_2^2 & \cdots & \lambda_2^k \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_n & \lambda_n^2 & \cdots & \lambda_n^k \end{bmatrix} \mathbf{c}^{(k)} \right) \right\|_2 \\
&= \sum_{i=1}^{n} |\boldsymbol{\beta}_i/\lambda_i|^2 |p_k(\lambda_i)|^2
\end{aligned} \tag{7.17}$$

So the convergence results for GMRES still hold for CG, but everywhere $\boldsymbol{\beta}_i/\lambda_i$ has to be substituted for $\boldsymbol{\beta}_i$. Note that this makes the weighing of the rows in Equation (7.17) decrease *less* than for GMRES. CG will probably still try to satisfy the rows corresponding to the largest eigenvalues, though. Remember that if a polynomial interpolates values close to the origin, it will have *huge* slopes far from the origin. This means that the polynomial would attain large values for the largest eigenvalues, which in turn would lead to a very large residual norm.



CHAPTER 8

Implementation Issues

implement, *to put into practical effect; carry out*
— THE AMERICAN HERITAGE DICTIONARY OF THE ENGLISH LANGUAGE

A particular method is not worth much unless it can be implemented efficiently and GMRES is no exception. The creators of GMRES, Saad and Schultz, showed in 1986 how to do it both elegantly and efficiently [SS86]. This chapter provides implementation details which uses many of their original ideas. In Appendix E.1 it is shown how GMRES can be implemented in MATLAB.

8.1 Implementation of GMRES

Let $\mathbf{A} \in \mathbb{C}^{n\times n}$ and $\mathbf{b} \in \mathbb{C}^{n}$. The approximate solution computed by GMRES after the kth iteration is then

$$\mathbf{x}^{(k)} = \underset{\mathbf{x}\in\mathbf{x}^{(0)}+\mathcal{K}_k(\mathbf{A},\mathbf{r}^{(0)})}{\operatorname{argmin}} \|\mathbf{b}-\mathbf{A}\mathbf{x}\|_2. \tag{8.1}$$

The idea behind GMRES is to represent the Krylov subspace in terms of orthogonal vectors so that

$$\mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)}) = \mathcal{R}(\mathbf{K}^{(k)}) = \mathcal{R}(\mathbf{V}^{(k)}), \tag{8.2}$$

for all $k \geq 1$ where

$$\mathbf{K}^{(k)} = \begin{bmatrix} \mathbf{r}^{(0)} & \mathbf{A}\mathbf{r}^{(0)} & \cdots & \mathbf{A}^{k-1}\mathbf{r}^{(0)} \end{bmatrix},$$

and the matrix

$$\mathbf{V}^{(k)} = \begin{bmatrix} \mathbf{v}^{(1)} & \mathbf{v}^{(2)} & \cdots & \mathbf{v}^{(k)} \end{bmatrix}$$

has orthonormal columns with $\mathbf{v}^{(1)} = \mathbf{r}^{(0)}/\|\mathbf{r}^{(0)}\|_2$.

We will assume that the columns of $\mathbf{K}^{(k)}$ are linearly independent so that $\mathbf{V}^{(k)}$ exists and can be produced by simple (modified) Gram–Schmidt orthonormalization. We now get

$$\begin{aligned} \mathbf{A}\mathbf{v}^{(k)} &= \mathbf{A}\mathbf{K}^{(k)}\mathbf{c}^{(k)} = \begin{bmatrix} \mathbf{r}^{(0)} & \mathbf{A}\mathbf{K}^{(k)} \end{bmatrix} \begin{bmatrix} 0 \\ \mathbf{c}^{(k)} \end{bmatrix} \\ &= \mathbf{K}^{(k+1)} \begin{bmatrix} 0 \\ \mathbf{c}^{(k)} \end{bmatrix} = \mathbf{V}^{(k+1)}\mathbf{y}^{(k)}, \end{aligned}$$

for some vectors $\mathbf{c}^{(k)} \in \mathbb{C}^k$ and $\mathbf{y}^{(k)} \in \mathbb{C}^{k+1}$. An equivalent way of writing this is

$$\mathbf{A}\mathbf{v}^{(k)} = \sum_{i=1}^{k+1} \mathbf{v}^{(i)} h_{i,k}. \tag{8.3}$$

Now assume that $\mathbf{v}^{(i)}$ is known for $i = 1, 2, \ldots, k$. Because of the orthonormality of these vectors we have

$$h_{i,k} = (\mathbf{A}\mathbf{v}^{(k)}, \mathbf{v}^{(i)})$$

for all $1 \leq i \leq k$. Isolating the last term in the summation in Equation (8.3) yields

$$\mathbf{v}^{(k+1)} h_{k+1,k} = \mathbf{A}\mathbf{v}^{(k)} - \sum_{i=1}^{k} \mathbf{v}^{(i)} h_{i,k} = \mathbf{s}^{(k)}, \tag{8.4}$$

which implies that $\mathbf{v}^{(k+1)} = \mathbf{s}^{(k)}/h_{k+1,k}$ where $h_{k+1,k} = \|\mathbf{s}^{(k)}\|_2$.

This way of producing an orthonormal sequence that spans a corresponding Krylov space is called the *Arnoldi process*. The relation can also be written in matrix form as

$$\mathbf{A}\mathbf{V}^{(k)} = \mathbf{V}^{(k+1)}\mathbf{H}^{(k)}, \tag{8.5}$$

where[1]

$$\mathbf{H}^{(k)} = \begin{bmatrix} h_{1,1} & h_{1,2} & \cdots & \cdots & h_{1,k} \\ h_{2,1} & h_{2,2} & \cdots & \cdots & h_{2,k} \\ 0 & \ddots & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \vdots \\ 0 & \cdots & \cdots & h_{k,k-1} & h_{k,k} \\ 0 & \cdots & \cdots & 0 & h_{k+1,k} \end{bmatrix} \in \mathbb{C}^{k+1 \times k}. \tag{8.6}$$

8.1.1 The Algorithm

Let us consider iteration k of GMRES and the residual $\mathbf{b} - \mathbf{A}\mathbf{x}$. As seen earlier, $\mathbf{x}$ must be of the form $\mathbf{x} \in \mathbf{x}^{(0)} + \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})$, or equivalently $\mathbf{x} = \mathbf{x}^{(0)} + \mathbf{V}^{(k)}\mathbf{y}$ for some $\mathbf{y} \in \mathbb{C}^k$. Using this last expression for $\mathbf{x}$ the residual becomes

$$\begin{aligned} \mathbf{b} - \mathbf{A}\mathbf{x} &= \mathbf{b} - \mathbf{A}\left(\mathbf{x}^{(0)} + \mathbf{V}^{(k)}\mathbf{y}\right) = \mathbf{r}^{(0)} - \mathbf{A}\mathbf{V}^{(k)}\mathbf{y} \\ &= \rho\mathbf{v}^{(1)} - \mathbf{V}^{(k+1)}\mathbf{H}^{(k)}\mathbf{y} = \mathbf{V}^{(k+1)}\left(\rho\mathbf{I}_1 - \mathbf{H}^{(k)}\mathbf{y}\right), \end{aligned}$$

where we have used the relation (8.5) and the fact that $\mathbf{r}^{(0)} = \rho\mathbf{v}^{(1)}$ with $\rho = \|\mathbf{r}^{(0)}\|_2$.

Since pre-multiplying with a matrix containing orthonormal columns does not change a 2-norm,[2] we get

$$\min_{\mathbf{x} \in \mathbf{x}^{(0)} + \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})} \|\mathbf{b} - \mathbf{A}\mathbf{x}\|_2 = \min_{\mathbf{y} \in \mathbb{C}^k} \left\|\rho\mathbf{I}_1 - \mathbf{H}^{(k)}\mathbf{y}\right\|_2. \tag{8.7}$$

[1] We use $h_{i,j}$ as the entries of $\mathbf{H}^{(k)}$ instead of the usual $\mathbf{H}^{(k)}_{i,j}$ since it makes the following algorithm and expressions simpler and more readable. Furthermore, the entries at i, j of $\mathbf{H}^{(k)}$ and $\mathbf{H}^{(k+1)}$ will be *identical*.

[2] When $\mathbf{Q}^H\mathbf{Q} = \mathbf{I}$ we have $\|\mathbf{Q}\mathbf{w}\|_2^2 = \mathbf{w}^H\mathbf{Q}^H\mathbf{Q}\mathbf{w} = \mathbf{w}^H\mathbf{w} = \|\mathbf{w}\|_2^2$.

Algorithm 1 Generalized Minimum Residual (GMRES)

```
1:  r^(0) ← b − Ax^(0)
2:  h_{1,0} ← ||r^(0)||_2
3:  k ← 0
4:  while h_{k+1,k} > 0 do
5:      k ← k + 1
6:      v^(k) ← r^(k−1) / h_{k,k−1}
7:      r^(k) ← Av^(k)
8:      for i = 1 to k do
9:          h_{i,k} ← (r^(k), v^(i))
10:         r^(k) ← r^(k) − h_{i,k} v^(i)
11:     end for
12:     h_{k+1,k} ← ||r^(k)||_2
13:     y^(k) ← argmin_y || h_{1,0} I_1 − H^(k) y ||
14:     x^(k) ← x^(0) + V^(k) y^(k)
15: end while
```

This is a simplified formulation of the original least-squares problem (8.1) and the Arnoldi process can finally be combined into the GMRES algorithm, seen as Algorithm 1. For each k the relation (8.6) holds and we have set $h_{1,0} = \rho = \|\mathbf{r}^{(0)}\|_2$.

8.1.2 In Case of Breakdown

Assume that we encounter $h_{k+1,k} = 0$ for the *first* time. This is often called *breakdown* of the algorithm. But it is not necessarily a bad thing. It means that

$$\mathbf{A}\mathbf{V}^{(k)} = \mathbf{V}^{(k)}\widehat{\mathbf{H}}^{(k)}, \tag{8.8}$$

where $\widehat{\mathbf{H}}^{(k)} \in \mathbb{C}^{k\times k}$ is equal to $\mathbf{H}^{(k)}$ (see Equation (8.6)) with the last row removed. The structure of $\widehat{\mathbf{H}}^{(k)}$ is called *upper Hessenberg*. From the above relation follows that

$$\mathbf{A}\mathbf{v}^{(k)} = \sum_{i=1}^{k} \mathbf{v}^{(i)} h_{i,k},$$

and since $\mathbf{v}^{(i)}$ is a linear combination of $\mathbf{r}^{(0)}, \mathbf{A}\mathbf{r}^{(0)}, \dots, \mathbf{A}^{k-1}\mathbf{r}^{(0)}$ for all i (cf. Equation (8.2)) we have that

$$\alpha_0\mathbf{r}^{(0)} + \alpha_1\mathbf{A}\mathbf{r}^{(0)} + \cdots + \alpha_k\mathbf{A}^k\mathbf{r}^{(0)} = \mathbf{0} \tag{8.9}$$

for some α_i, that is, the vectors are linearly dependent. Note furthermore that we must have $\alpha_k \neq 0$ since we otherwise would have had an algorithm breakdown earlier (see Equation (8.4)).

But linear dependency is exactly what we would like. In Section 7.3 we saw that linear dependency in the Krylov subspace was a necessary condition for a Krylov solution to exist.

If $\alpha_0 = 0$ in Equation (8.9) we see that

$$\mathbf{A}\left(\alpha_1\mathbf{r}^{(0)} + \alpha_2\mathbf{A}\mathbf{r}^{(0)} + \cdots + \alpha_k\mathbf{A}^{k-1}\mathbf{r}^{(0)}\right) = \mathbf{A}\mathbf{w} = 0.$$

Since $\mathbf{w} \in \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})$ we have from Theorem 7.7 that $\mathbf{A}$ is singular and no Krylov solution exists.

If we have $\alpha_0 \neq 0$ we get

$$\begin{aligned}\mathbf{r}^{(0)} &= \mathbf{A}\left[-\frac{1}{\alpha_0}\left(\alpha_1\mathbf{r}^{(0)} + \alpha_2\mathbf{A}\mathbf{r}^{(0)} + \cdots + \alpha_k\mathbf{A}^{k-1}\mathbf{r}^{(0)}\right)\right] \\ &= \mathbf{A}(\mathbf{x}^{(k)} - \mathbf{x}^{(0)})\end{aligned}$$

so a solution does exists. Notice that since $\alpha_0 = 0$ implied that $\mathbf{A}$ is singular, we must have that a regular $\mathbf{A}$ implies $\alpha_0 \neq 0$. This means: When a Krylov method stops because of linearly dependency and $\mathbf{A}$ is regular, a solution will be found.

Notice that linearly dependency, or equivalently $h_{k+1,k} = 0$, will happen for $k = m$ at the latest, where m is the degree of the minimal polynomial, see Theorem 7.2.

8.1.3 Rewriting the Least-Squares Problem

How should the least-squares problem, stated in (8.7) and in line 13 of the algorithm, be solved efficiently? A good way to do this is to split

$\mathbf{H}^{(k)}$ into the product of an orthogonal matrix and an upper triangular matrix:[3]

$$\mathbf{H}^{(k)} = \mathbf{Q}^{(k)}\mathbf{T}^{(k)}. \tag{8.10}$$

Since $\mathbf{H}^{(k)} \in \mathbb{C}^{(k+1)\times k}$ we have that $\mathbf{Q}^{(k)} \in \mathbb{C}^{(k+1)\times(k+1)}$ and $\mathbf{T}^{(k)} \in \mathbb{C}^{(k+1)\times k}$. From iteration $k-1$ to the next, a zero row and a column must be appended to $\mathbf{H}^{(k-1)}$ in order to form $\mathbf{H}^{(k)}$, so we would like to be able to exploit the factorization already obtained. Let the last column of $\mathbf{H}^{(k)}$ be $\begin{bmatrix}\mathbf{h}\\ \eta\end{bmatrix}$ and consider

$$\begin{aligned}
\mathbf{H}^{(k)} &= \begin{bmatrix}\mathbf{H}^{(k-1)} & \mathbf{h}\\ \mathbf{0}^H & \eta\end{bmatrix} = \begin{bmatrix}\mathbf{Q}^{(k-1)}\mathbf{T}^{(k-1)} & \mathbf{h}\\ \mathbf{0}^H & \eta\end{bmatrix}\\
&= \begin{bmatrix}\mathbf{Q}^{(k-1)} & \mathbf{0}\\ \mathbf{0}^H & 1\end{bmatrix}\begin{bmatrix}\mathbf{T}^{(k-1)} & {\mathbf{Q}^{(k-1)}}^H\mathbf{h}\\ \mathbf{0}^H & \eta\end{bmatrix}\\
&= \left(\begin{bmatrix}\mathbf{Q}^{(k-1)} & \mathbf{0}\\ \mathbf{0}^H & 1\end{bmatrix}\mathbf{G}\right)\left(\mathbf{G}^H\begin{bmatrix}\mathbf{T}^{(k-1)} & {\mathbf{Q}^{(k-1)}}^H\mathbf{h}\\ \mathbf{0}^H & \eta\end{bmatrix}\right)
\end{aligned} \tag{8.11}$$

which is valid whenever $\mathbf{G}\mathbf{G}^H = \mathbf{I}$.

Now if $\mathbf{G} = \mathbf{I}$, we see that $\mathbf{H}^{(k)}$ is written as the product of two matrices where the first is orthogonal and the second is nearly upper triangular, only η, the element at $(k+1,k)$, spoils this. If we can choose an orthogonal $\mathbf{G}$ so that the last factor becomes upper triangular, we have a new valid factorization. This is obtained by using a *Givens rotation*. In this case, the matrix must be

$$\mathbf{G} = \begin{bmatrix}\mathbf{I} & & \\ & c & s\\ & -s & c\end{bmatrix} \in \mathbb{C}^{(k+1)\times(k+1)},$$

[3]This factorization is typically called QR-factorization, where Q represents the orthogonal matrix and R the triangular matrix. To avoid confusion with residual vectors, we use the letter T for the triangular matrix.

where $c = \cos(\theta)$ and $s = \sin(\theta)$ for the same θ, which ensures that $\mathbf{G}$ is orthogonal. We wish to find c and s so that

$$\mathbf{G}^H\begin{bmatrix}{\mathbf{Q}^{(k-1)}}^H\mathbf{h}\\ \eta\end{bmatrix} = \begin{bmatrix}\mathbf{I} & & \\ & \overline{c} & -\overline{s}\\ & \overline{s} & \overline{c}\end{bmatrix}\begin{bmatrix}{\mathbf{Q}^{(k-1)}}^H\mathbf{h}\\ \eta\end{bmatrix} = \begin{bmatrix}\mathbf{k}\\ 0\end{bmatrix}$$

If the bottom element of ${\mathbf{Q}^{(k-1)}}^H\mathbf{h}$ is denoted γ, straightforward calculations show that[4]

$$s = \frac{1}{\sqrt{1+|\gamma/\eta|^2}} \quad\text{and}\quad c = -\frac{\gamma/\eta}{\sqrt{1+|\gamma/\eta|^2}}.$$

With this $\mathbf{G}$, the two set of parentheses in Equation (8.11) will represent $\mathbf{Q}^{(k)}$ and $\mathbf{T}^{(k)}$ respectively.

The norm in Equation (8.7) can now be rewritten as

$$\left\|\rho\mathbf{I}_1 - \mathbf{H}^{(k)}\mathbf{y}\right\|_2 = \left\|\rho\mathbf{I}_1 - \mathbf{Q}^{(k)}\mathbf{T}^{(k)}\mathbf{y}\right\|_2 = \left\|\rho{\mathbf{Q}^{(k)}}^H\mathbf{I}_1 - \mathbf{T}^{(k)}\mathbf{y}\right\|_2. \tag{8.12}$$

Note, however, that the last row of $\mathbf{T}^{(k)} \in \mathbb{C}^{(k+1)\times k}$ is *always* zero. This makes it impossible to fulfill the last row in the least-squares problem.

Let $\widehat{\mathbf{T}^{(k)}}$ represent $\mathbf{T}^{(k)}$ without the last row and likewise for $\widehat{{\mathbf{Q}^{(k)}}^H\mathbf{I}_1}$. If $\widehat{\mathbf{T}^{(k)}}$ has non-negative elements along the diagonal, the inverse exists and the minimum of the norm in (8.12) can be found explicitly. This means that

$$\mathbf{y}^{(k)} = \underset{\mathbf{y}}{\operatorname{argmin}}\left\|\rho\widehat{{\mathbf{Q}^{(k)}}^H}\mathbf{I}_1 - \widehat{\mathbf{T}^{(k)}}\mathbf{y}\right\|_2 = \rho\widehat{\mathbf{T}^{(k)}}^{-1}\widehat{{\mathbf{Q}^{(k)}}^H}\mathbf{I}_1$$

and the solution can be expressed as:

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + \mathbf{V}^{(k)}\mathbf{y}^{(k)} = \mathbf{x}^{(0)} + \rho\mathbf{V}^{(k)}\widehat{\mathbf{T}^{(k)}}^{-1}\widehat{{\mathbf{Q}^{(k)}}^H}\mathbf{I}_1.$$

[4]If $\eta = 0$ we just choose $\mathbf{G} = \mathbf{I}$.

Notice that this implies that all rows in Equation (8.12) will be fulfilled except for the last. So the residual norm is known explicitly to be the absolute value of the bottom element of $\rho \mathbf{Q}^{(k)^H} \mathbf{I}_1$. This means that one knows the residual norm in each iteration at no extra cost, without having to compute $\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|_2$.

Assume that the newly found element at (k, k) of $\mathbf{T}^{(k)}$ is zero. Then $\widehat{\mathbf{T}^{(k)}}$ is singular and the above formula cannot be used. It is easily shown that the absolute value of this corner element is $\sqrt{|\eta|^2 + |\gamma|^2}$, so $\mathbf{T}^{(k)}_{k,k} = 0$ implies $\eta = \gamma = 0$. But when $\eta = h_{k+1,k} = 0$ we have linearly dependency in the Krylov subspace, and equations (8.8) and (8.10) yield

$$\mathbf{A}\mathbf{V}^{(k)} = \mathbf{V}^{(k)}\widehat{\mathbf{H}}^{(k)} = \mathbf{V}^{(k)}\mathbf{Q}^{(k)}\widehat{\mathbf{T}^{(k)}}.$$

Since $\widehat{\mathbf{T}^{(k)}}$ is singular we can find a vector $\mathbf{y}$ such that $\widehat{\mathbf{T}^{(k)}}\mathbf{y} = \mathbf{0}$. That makes the right-hand side of the above expression equal to zero, so we also have $\mathbf{A}\mathbf{V}^{(k)}\mathbf{y} = \mathbf{0}$. Since $\mathbf{V}^{(k)}\mathbf{y} \in \mathcal{K}_k(\mathbf{A}, \mathbf{r}^{(0)})$ we have from Theorem 7.7 that $\mathbf{A}$ is singular and no Krylov solution exists. So this way of implementing GMRES can actually detect when a solution can not be found.

Now assume that the (k, k)–element of $\widehat{\mathbf{T}^{(k)}}$ is nonzero, which means that it is invertible. We introduce the auxiliary matrix,

$$\mathbf{W}^{(k)} = \begin{bmatrix}\mathbf{w}^{(1)} & \mathbf{w}^{(2)} & \cdots & \mathbf{w}^{(k)}\end{bmatrix} = \mathbf{V}^{(k)}\widehat{\mathbf{T}^{(k)}}^{-1},$$

for all k. From this follows that[5]

$$\begin{aligned}
\mathbf{v}^{(k)} &= \mathbf{W}^{(k)}\mathbf{t}^{(k)} = \begin{bmatrix}\mathbf{W}^{(k-1)} & \mathbf{w}^{(k)}\end{bmatrix}\begin{bmatrix}\mathbf{t}^{(k)}_{1:k-1} \\ \mathbf{t}^{(k)}_k\end{bmatrix} \\
&= \mathbf{W}^{(k-1)}\mathbf{t}^{(k)}_{1:k-1} + \mathbf{w}^{(k)}\mathbf{t}^{(k)}_k \quad \Rightarrow \\
\mathbf{w}^{(k)} &= \left(\mathbf{v}^{(k)} - \mathbf{W}^{(k-1)}\mathbf{t}^{(k)}_{1:k-1}\right)/\mathbf{t}^{(k)}_k
\end{aligned}$$

for $k > 1$. When $k = 1$ we just have $\mathbf{w}^{(1)} = \mathbf{v}^{(1)}/\mathbf{t}^{(1)}_1$.

[5]Here, we also use $\widehat{\mathbf{T}^{(k)}} = \begin{bmatrix}\mathbf{t}^{(1)} & \mathbf{t}^{(2)} & \cdots & \mathbf{t}^{(k)}\end{bmatrix}$.

By updating $\mathbf{W}^{(k)}$ in this fashion, the approximate solution at each step can be found as

$$\mathbf{x}^{(k)} = \mathbf{x}^{(0)} + \rho\mathbf{V}^{(k)}\widehat{\mathbf{T}^{(k)}}^{-1}\widehat{\mathbf{Q}^{(k)^H}}\mathbf{I}_1 = \mathbf{x}^{(0)} + \rho\mathbf{W}^{(k)}\widehat{\mathbf{Q}^{(k)^H}}\mathbf{I}_1.$$

In Appendix E.1, MATLAB code can be found illustrating all the implementation aspects of the GMRES method mentioned in this section.

Note that MATLAB 6.0's own implementation of GMRES does *not* take special actions when the corner element of $\widehat{\mathbf{T}^{(k)}}$ becomes zero. This means that in some cases, their algorithm outputs an ugly `divide by zero` warning, although it warns the user that no solution has been found. See Section D.6 for an example of this.

8.2 Related Algorithms and Practical Considerations

The fact that full orthonormalization has to be done in each step of GMRES makes the work per iteration, and overall storage requirements, increase. To get around this, the algorithm *restarted* GMRES is often used instead. Given a starting guess, GMRES is run for k iterations. Then the approximate solution found is used as a starting guess to start up GMRES once again. And so forth. Since GMRES this way is restarted every k iterations, it is also called GMRES(k). Convergence analysis of this algorithm becomes much harder. Important is it to note however, that when it comes to discrete ill-posed problems, GMRES has a tendency to converge very quickly or not at all. So restarted GMRES may not be a relevant method for these kinds of problems.

Another thing that is important for practical use is stopping criteria. We know that the residual always is non-increasing, but semi-convergence, which is often present for discrete ill-posed problems, shows that one should not always choose the solution with the smallest residual norm. A concept introduced by P. C. Hansen (see [Han98a]) is the *L-curve*. This is a way of choosing the best regularization parameter (which for iterative methods typically is the iteration step) by considering the two quantities $\|\mathbf{x}^{(k)}\|_2$ and $\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|_2$. The residual norm will typically decrease quickly in the beginning, then level off and then accelerate the decrease once again. The solution norm typically increases in the beginning, then levels off and then increases rapidly as noise begins to influence the solution. This phenomenon is often seen as a big "L" when plotting $\log\|\mathbf{x}^{(k)}\|_2$ and $\log\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|_2$ against each other, hence the name. The trick is now to pick the solution that corresponds to the corner of the L, since this can be shown to be a good choice of regularization parameter in many cases. The down-side to this scheme, in relation to GMRES, is that the solution norm has to be computed in each iteration step, which is not needed in the traditional implementation.

A related, but different approach, is taken in the article [CLR00]. Here, they look at L-curves made by the quantities $\|\kappa^{(k)}\|_2$ and $\|\mathbf{b} - \mathbf{A}\mathbf{x}^{(k)}\|_2$, where $\kappa^{(k)}$ is the condition number of the matrix $\mathbf{H}^{(k)}$ (see equations (3.11) and (8.7)) used to solve the smaller, reduced linear system. They show how this method often is a better approach than the traditional L-curve.

There is no shortage of Krylov subspace methods. Over 20 of them popped up when counting loosely in the literature. Methods like GCR (Generalized Minimal Residual), ORTHOMIN, GENCR and ORTHODIR are all mathematically equivalent to GMRES, but differ in the implementation. As already mentioned, MINRES is also mathematically equivalent to GMRES, but is tailor-made for symmetric coefficient matrices.

Quasi-optimal Krylov subspace methods are methods that generate non-optimal solutions. A number of Quasi GMRES-related methods include Bi-CG, QMR (Quasi Minimal Residual), CG-S, Bi-CG-STAB, restarted GMRES and hybrid GMRES. See [Gre97] or [SvdV93] for a description for most of these algorithms.

Why make two different methods that are mathematically equivalent? There are at least two reasons for this: To improve efficiency and/or accuracy. Efficiency is basically the time it takes for the method to find a satisfactory solution. Memory usage is also an important factor here.

Practical implementations are always done on finite precision computers, and the effect of round-off errors is a delicate matter. Methods, such as GMRES, that rely on orthogonalizing the Krylov subspaces, can exhibit unwanted behavior if orthogonality is not exactly fulfilled due to rounding errors. Both [Kar91] or [Gre97] touch upon these subjects.

Chapter 9

Conclusion

conclusion, *the result or outcome of an act or process*
— The American Heritage Dictionary of the English Language

The question that set off this project was: *"How well does GMRES work for discrete linear ill-posed problems?"* It was decided to attack this question from an abstract point of view, trying to generalize conclusions/properties as much as possible before diving into specifics. And instead of attacking the question posed above head-on we would go back and see where discrete ill-posed problems come from and what can be learned from these. Note that this was not necessary in order to answer the above question, but it was an interesting area that fitted well into the spirit of generalizing.

The adjective *ill-posed* can be used for a huge number of mathematical problems. If a unique solution to a certain problem always exists and the solution process is stable the problem is called well-posed, otherwise ill-posed. In this project the class of problems considered where narrowed down to $Kf = g$ where K was a compact linear operator. This class was very relevant since it includes all finite matrix mappings and nearly all integral operators.

In order to analyze GMRES, it quickly became clear that the eigenvalues of the coefficient matrix were an essential factor. So the strategy was to analyze the behavior of the eigenvalues (-vectors) of compact linear operators, to find out how these quantities carried over to the finite dimensional matrix approximations, and finally to analyze GMRES using the acquired knowledge.

In the analysis of compact linear operators, focus was put on both eigenvalues and singular values of such. Singular values are a popular tool for analyzing linear ill-posed problems and any knowledge of these could maybe lead to increased understanding of the eigenvalues.

Singular values are generally much more well-behaved than eigenvalues since they are eigenvalues of self-adjoint, non-negative definite operators of the form K^*K. This means that the singular values are always non-negative and that orthonormal vectors always exist that span the range of the operator. For eigenvalues, in general, this is not so. The number of eigenvalues may be none, finite or infinite and the corresponding eigenvectors need not be orthonormal. But eigenvalues are always denumerable and when there are an infinite number, they tend towards zero.

The existence of eigenvalues is a subtle matter. For instance, Volterra integral operators can only have zero as eigenvalues, if any at all. When an operator is normal ($K^*K = KK^*$) eigenvalues always exist such that the corresponding eigenvectors are orthonormal and span the range of the operator. In this case, the absolute value of the eigenvalues and the singular values are identical. When an operator is degenerate, i.e. a finite sum of one-dimensional mappings, the number of non-zero eigenvalues will be equal to or less than the dimension of the range.

The asymptotic behavior of eigenvalues and singular values was important to investigate in this project. A comprehensive article on the eigenvalues of integral operators was published by Hille and Tamarkin in 1931. In it, they derived a number of theorems related to different integral kernel properties. In the years thereafter some of the bounds were improved, but this article still seems like one of the most important contributions to this area. To roughly summarize the results: When an integral kernel is analytic (a subset of C^∞) the decay of the eigenvalues will be at least exponential. When the kernel is only a finite number of times continuously differentiable, the decay will be polynomial.

The discretization technique considered in this thesis was Galerkin discretization. This was done by projecting vectors in the domain and range onto finite dimensional subspaces spanned by orthogonal sets of vectors. For singular values and -vectors, it was found that these quantities always converge to the true ones as the discretization size increased. Furthermore, error bounds were derived.

The results were similar for approximate eigenvalues and -vectors. The most important difference however, was that given a sequence of approximate eigenvalues converging to some element, the limit would be *either* zero or an eigenvalue of the operator, which we were trying to approximate.

The focus could now be put on GMRES or, in the extent possible, on Krylov subspace methods in general. Krylov subspace methods could be defined as methods where the solution in iteration k was found in a Krylov subspace, the span of $\mathbf{A}^i$, $i = 0, 1, \ldots, k-1$ applied to some (constant) vector.

The solution to $\mathbf{A}\mathbf{x} = \mathbf{b}$ would always lie in such a Krylov subspace, unless for some cases when zero was an eigenvalue of $\mathbf{A}$ with index larger than 1 (these cases were possible to detect easily in an implementation of GMRES). Furthermore the solution, if it existed, would always be found in a maximum of n iterations ($\mathbf{A} \in \mathbb{C}^{n\times n}$).

When $\mathbf{A}$ was assumed diagonalizable it was possible to do a convergence analysis for GMRES. Assuming that the eigenvector components of the right-hand side decayed sufficiently fast in roughly the same order as the eigenvalues decayed, it was possible to derive a good approximate bound for the residual norm. The strategy of GMRES in the kth iteration seemed to be to build the approximate solutions from (roughly) the k eigenvectors corresponding to the largest eigenvalues. If the right-hand side eigenvector components from some point on stopped decaying fast, e.g. due to noise, the bound would still be valid although somewhat pessimistic.

Now returning to the initial question of how well GMRES works for discrete ill-posed problems. An easy answer: It *can* work very well. As can be seen from the examples in the appendix, it can go either way. It would require thorough comparison to *other* algorithms to be able to comment on *how* good this algorithm is.

Some conditions must be met in order for GMRES to be successful. The eigenvector components of the right-hand side must decrease *at least* as quickly as the eigenvalues. If they do not, GMRES will tend to construct a solution from the eigenvectors corresponding to small eigenvalues. This is unfortunate since these eigenvectors often are highly oscillatory[1] and can lead to unwanted looking solutions. This condition can be seen as a kind of Picard condition for the eigenvalues, although it needs to be more precisely defined for practical use.

If the right-hand side eigenvector components decrease sufficiently fast but level out from some point, then this suggests that the right-hand side is influenced by noise. Furthermore, GMRES should be stopped after a number of iterations corresponding to where the right-hand side eigenvector components begin to level out. Unfortunately, the quantities needed to make such a decision are not known in practise. Instead one could turn to L-curve stopping criteria that detects when the solution begins to "explode" due to noise.

Even if the right-hand side eigenvector components decrease sufficiently fast, one must still be cautious. This just implies that the *residual norm* will decrease approximately equally quickly. But there is no guarantee for the error norm. Only if the eigenvectors appear to be able to assemble a satisfactory solution will it be likely that GMRES will succeed.

Some remarks can be said about the regularization properties of Krylov subspace methods in general. Since the image of integral operators (almost) always will be smooth, a matrix discretization will have similar properties. This means that every Krylov subspace, apart from possibly the starting vector, will be spanned by smooth vectors. Since

[1]That the eigenvectors become more and more oscillatory as the corresponding eigenvalues decrease is merely an empirical result. It can be shown for the ETP class, though, see Section 4.4.3.

the solutions always lie in such a Krylov subspace, the solutions will be similarly smooth.

9.1 The Need for Further Investigation

Some areas still need to be investigated in order to more fully understand the behavior of GMRES for discrete ill-posed problems. Some of these areas are listed below.

- The effects of finite precision and noisy data. All results in this thesis were derived assuming infinite precision. More experiments concerning noisy right-hand sides need to be made, and theoretical (statistical) models of such problems can perhaps provide better convergence results.
- Comparison to other methods. When asked *how* well GMRES works for certain problems, one ideally needs to compare to the best available algorithm to this exact kind of problem (the "goodness" of approximate solutions can seldom be globally measured, though). Since many algorithms are made to solve similar kinds of problems, comparisons need to be made to reveal the strengths and weaknesses of each.
- Eigenvectors of integral operators and their discretized versions. The oscillatory properties of eigenvectors are very relevant concerning regularization properties. Furthermore, since Krylov subspace methods tend to build their solutions from eigenvectors, any knowledge of their appearance is relevant.
- More can be learned from the derivation and physical meaning of the operator. Only focusing on analytical properties of kernels may hide important information. For instance, the physical "translation" of a given eigenvalue problem may provide insight not obtainable otherwise.
- The choice of basis. How crucial is the choice of basis used for Galerkin discretization concerning operator approximation and

the convergence of e.g. GMRES? Can a basis be chosen that makes the eigenvectors of $\mathbf{A}$ more fit to "build" a good solution?

- Results regarding other discretization schemes. In this thesis, only Galerkin discretization has been considered. Results on eigenvalue/-vector approximation is needed when discretizing differently.

- The starting guess for GMRES. How much does the starting guess influence the convergence of GMRES? Answering this may also provide insight into the convergence of restarted GMRES.

appendix (əpen'diks), *n.* (*pl.* **-dixes, -dices** (-sēz)) something appended; an adjunct or concomitant; a supplement to a book or document containing useful material; a small process arising from, or the prolongation of, any organ, esp. the vermiform appendix of the intestine.

APPENDIX A

Orthonormal Bases in L^2

We will here present orthogonal bases[1] for $L^2([a,b])$, $L^2([0,\infty[)$ and $L^2(\mathbb{R})$.

The bases presented on bounded intervals $[a,b]$ will have fixed values of a and b, but proper translation and scaling can easily provide a basis for another interval. For instance let an orthogonal basis $(e_n) \subset L^2([a,b])$ be given. If one wishes a basis $(\tilde{e}_n) \subset L^2([c,d])$ straightforward calculation yields:

$$\tilde{e}_n(t) = \sqrt{\frac{b-a}{d-c}} e_n\left((x-a)\frac{d-c}{b-a} + c\right).$$

Similar transformations can be made for unbounded intervals. Unless stated otherwise, the bases presented will be for *real* functions only.

[1]Some will only be orthogonal sequences.

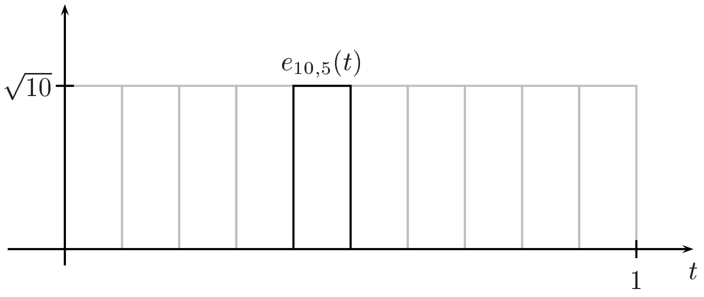


Figure A.1: *An orthonormal box function sequence.*

A.1 Box Functions

An orthonormal sequence of N box functions in $L^2([0,1])$ is simply

$$e_{N,i}(t) = \begin{cases} \sqrt{N} & \text{for } \frac{i-1}{N} \leq t \leq \frac{i}{N} \\ 0 & \text{otherwise} \end{cases}, \quad i = 1, 2, \dots, N.$$

If discretizing an integral operator K with kernel $k \in L^2([0,1]^2)$, the needed computation becomes

$$\begin{aligned} \mathbf{A}_{i,j} = (Ke_j, e_i) &= \int_0^1 \int_0^1 k(s,t) e_j(t) e_i(s) dt ds \\ &= \sqrt{MN} \int_{\frac{i-1}{N}}^{\frac{i}{N}} \int_{\frac{j-1}{M}}^{\frac{j}{M}} k(s,t) dt ds. \end{aligned} \tag{A.1}$$

A.2 Polynomials

We will first look at a basis for $L^2(I)$ consisting of polynomials on a closed and bounded interval I. The functions $1, t, t^2, \dots$ form a basis for the set of polynomials on I. Since polynomials are dense in $C(I)$ (by Weierstrass' approximation theorem), and since $C(I)$ is dense in

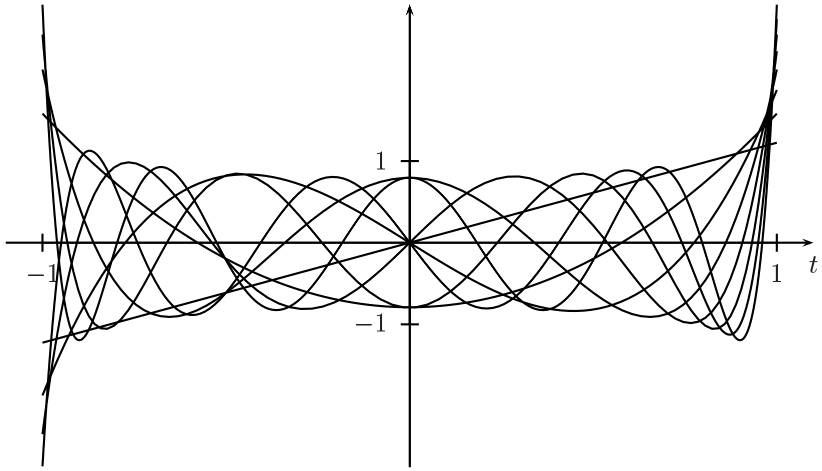


Figure A.2: *The (normalized) Legendre polynomials $P_1, P_2, \dots, P_8$.*

$L^2(I)$ *by definition*, these functions span $L^2(I)$. But the are not orthogonal. This can be accomplished by using the Gram–Schmidt orthonormalization technique, which leads to the following basis:

$$e_n(t) = \sqrt{\frac{2n+1}{2}} P_n(t), \quad n = 0, 1, 2, \dots,$$

where P_n are the *Legendre* polynomials. Some of these polynomials can be seen in Figure A.2.

When polynomials are orthogonal, a three-recursive formula can *always* be set up to define their coefficients (Theorem 3.2.1 in [Sze67]). For Legendre polynomials it is

$$\begin{aligned} &P_0(t) = 1, \quad P_1(t) = t, \\ &(n+1)P_{n+1}(t) = (2n+1)tP_n(t) - nP_{n-1}(t), \quad n = 1, 2, \dots. \end{aligned}$$

We will now turn to a basis, based on polynomials, for $L^2([0,\infty[)$. Since the *Laguerre* polynomials (L_n) are orthogonal with respect to a

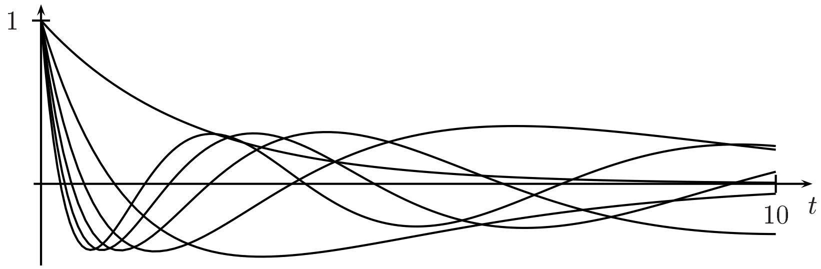


Figure A.3: *Orthonormal functions on $L^2([0,\infty[)$, based on Laguerre polynomials.*

different inner product than usual, namely

$$\int_0^\infty e^{-t} L_i(t) L_j(t) dt = \delta_{ij},$$

we see that we can obtain an orthonormal basis by using the functions

$$e_n(t) = e^{-\frac{t}{2}} L_n(t), \quad n = 0, 1, 2, \ldots.$$

A plot of some of these functions can be seen in Figure A.3. A three-recursive formula can also be set up for L_n in this case:

$$L_0(t) = 1, \quad L_1(t) = 1 - t,$$
$$(n+1)L_{n+1}(t) = (2n + 1 - t)L_n(t) - nL_{n-1}(t), \quad n = 1, 2, \ldots.$$

Finally we seek functions, based on polynomials, that are orthonormal in $L^2(\mathbb{R})$. Here the *Hermite* polynomials can be used. They have the following orthogonality:

$$\int_{\mathbb{R}} e^{-t^2} H_i(t) H_j(t) dt = \delta_{ij} \sqrt{\pi} 2^n n!.$$

This leads to the following functions which are orthonormal with respect to the usual L^2 inner product:

$$e_n(t) = \frac{e^{-\frac{t^2}{2}}}{\sqrt{\sqrt{\pi} 2^n n!}} H_n(t), \quad n = 0, 1, 2, \ldots.$$

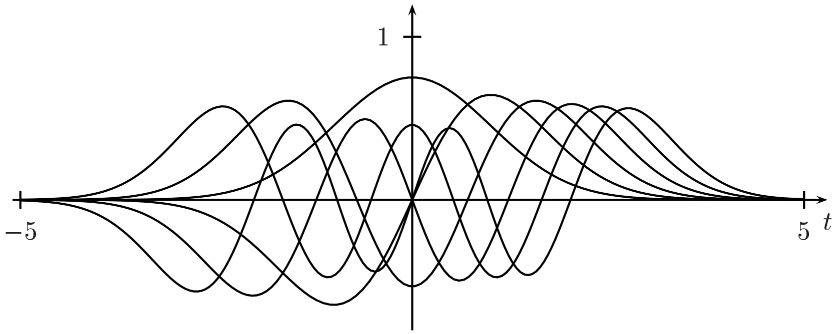


Figure A.4: *Orthonormal functions on $L^2(\mathbb{R})$, based on Hermite polynomials.*

Some of these are plotted in Figure A.4. The three-recursive formula here is:

$$H_0(t) = 1, \quad H_1(t) = 2t,$$
$$H_{n+1}(t) = 2tH_n(t) - 2nH_{n-1}(t), \quad n = 1, 2, \ldots$$

For more information on these polynomials and orthogonal polynomials in general, see [Ped00, p. 42] or [Sze67].

A.3 Trigonometric Basis Functions

We will here restrict ourselves to the interval $[0, \pi]$. A basis using cosine functions is

$$e_0(t) = \frac{1}{\sqrt{\pi}}, \quad e_n(t) = \sqrt{\frac{2}{\pi}} \cos(nt), \quad n = 1, 2, \ldots.$$

A plot of some of these functions can be seen in Figure A.5. Sine functions can also be used:

$$e_n(t) = \sqrt{\frac{2}{\pi}} \sin(nt), \quad n = 1, 2, \ldots.$$

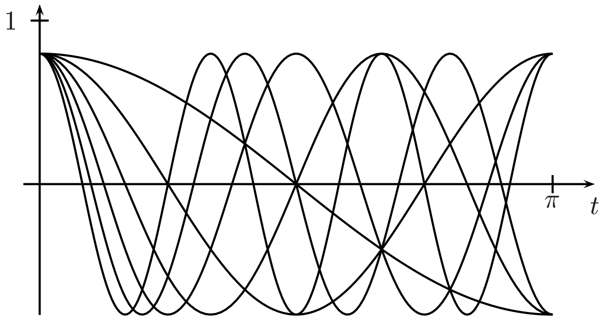


Figure A.5: *An orthonormal cosine sequence.*

Finally, to obtain a basis for the *complex* $L^2([-\pi,\pi])$ we can use

$$e_n(t) = \frac{1}{\sqrt{2\pi}} e^{int}, \quad n \in \mathbb{Z}.$$

A.4 A Simple Wavelet Basis

The perhaps simplest wavelet basis possible on $L^2([0,1])$ is the so-called *Haar* basis. This is easily defined using double indices. Let $h_0^1 = 1$ and

$$h_n^j(x) = \begin{cases} \sqrt{2^n}, & \text{if } x \in \left]\frac{j-1}{2^n}, \frac{j-\frac{1}{2}}{2^n}\right] \\ -\sqrt{2^n}, & \text{if } x \in \left]\frac{j-\frac{1}{2}}{2^n}, \frac{j}{2^n}\right] \\ 0, & \text{otherwise,} \end{cases}$$

for $n = 0, 1, 2, \ldots$ and $j = 1, 2, \ldots, 2^n$.

A.5 Orthonormal Bases in $L^2(I \times J)$

Let (e_m^I) and (e_n^J) be an orthonormal basis for $L^2(I)$ and $L^2(J)$ respectively. We now wish to construct an orthonormal basis for $L^2(I \times J)$ using these two bases.

Before proceeding, we must introduce the *tensor product*. Given $f \in L^2(I)$ and $g \in L^2(J)$ we define the function $f \otimes g$ by

$$f \otimes g(x,y) = f(x)g(y) \quad \text{for all } (x,y) \in I \times J. \tag{A.2}$$

Since

$$\int_I \int_J |f \otimes g(x,y)|^2 \, dydx = \int_I |f(x)|^2 dx \int_J |g(y)|^2 dy < \infty,$$

we have that $f \otimes g \in L^2(I \times J)$. Consider now the sequence $(e_m^I \otimes \overline{e_n^J})$. This sequence is orthonormal since

$$\begin{aligned} \left(e_m^I \otimes \overline{e_n^J}, e_{m'}^I \otimes \overline{e_{n'}^J}\right) &= \int_I \int_J e_m^I(s)\overline{e_n^J(t)}\,\overline{e_{m'}^I(s)\overline{e_{n'}^J(t)}}\,dtds \\ &= \int_I e_m^I(s)\overline{e_{m'}^I(s)}ds \int_J e_{n'}^J(t)\overline{e_n^J(t)}dt \\ &= \left(e_m^I, e_{m'}^I\right)\left(e_{n'}^J, e_n^J\right) = \delta_{mm'}\delta_{n'n}. \end{aligned}$$

To show it is a basis let $k \in L^2(I \times J)$ and assume

$$(k, e_m^I \otimes \overline{e_n^J}) = 0 \quad \text{for all } n, m.$$

Then, for all n and m we have

$$\begin{aligned} 0 &= \int_I \int_J k(s,t)\overline{e_m^I(s)\overline{e_n^J(t)}}\,dtds \\ &= \int_I \left(\int_J k(s,t)e_n^J(t)dt\right)\overline{e_m^I(s)}ds = (Ke_n^J, e_m^I), \end{aligned}$$

where K is the integral operator induced by the kernel k. Since the above expression holds for all m we must have $Ke_n^J = 0$ for all n, which in turn implies $K = 0$. Since an integral operator is uniquely determined by its kernel we have $k = 0$ which shows that the sequence $(e_m^I \otimes \overline{e_n^J})$ is total, and hence is an orthonormal basis.

APPENDIX B

Special Operators

As mentioned in Section 4.1, a large part of integral operators are compact. But the set of compact operators is *not* a subset of integral operators and vice versa. The following two examples will illustrate this fact.[1]

B.1 A Compact Operator not Integral

First we prove a *necessary* condition for an operator to be integral.

Theorem B.1 *Set $I = [0,1]$ and let $K : L^2(I) \to L^2(I)$ be an integral operator with kernel $k : I \times I \to \mathbb{C}$. Then there exists a function $M : I \to [0, \infty[$ such that*

$$f \in L^\infty(I) \quad \Rightarrow \quad |Kf(s)| \leq \|f\|_\infty M(s) \quad \textit{almost everywhere.}$$

Proof: Since $1_I \in L^2(I)$ we have (from the definition of the domain, see Definition (4.2))

$$M(s) = \int_I |k(s,t)\, 1| dt < \infty,$$

[1] The examples are adapted from [HS78].

which at the same time defines M explicitly. Now we have for $f \in L^\infty(I)$,

$$|Kf(s)| = \left|\int_I k(s,t)f(t)dt\right| \le \|f\|_\infty \int_I |k(s,t)|dt < \|f\|_\infty M(s).$$

□

Now consider the cosine basis on I, denote it (e_n), and the Haar basis on I, denote it (h_n). Note that single index is used for the Haar basis. This can easily be done by ordering the functions as

$$(h_0^0; h_1^0, h_1^1; h_2^0, h_2^1, h_2^2, h_2^3; \ldots).$$

Define now an operator L by

$$Lx = \sum_{n\in\mathbb{N}} \mu_n(x, e_n)h_n$$

where $\mu_n = 1/\sqrt{\|h_n\|_\infty}$. L is obviously linear. Since the sequence (μ_n) is bounded and $\mu_n \to 0$ for $n \to \infty$ the operator is bounded and compact (see Theorem 2.17). We have $e_n \in L^\infty(I)$ and consider

$$|Le_n(s)| = |\mu_n h_n(s)|,$$

which can be made arbitrarily large for almost all $s \in I$. This means that a function M can not be found so

$$|Le_n(s)| \le \|e_n\|_\infty M(s)$$

almost everywhere. Hence, L can not be an integral operator.

B.2 An Integral Operator not Compact

Let $I = J = [0, \infty[$ and define the kernel of the integral operator K as

$$k(s,t) = \begin{cases} 1 & \text{if } n-1 \le s,t \le n \text{ for some } n \in \mathbb{N} \\ 0 & \text{otherwise} \end{cases}$$

This kernel does obviously not lie in $L^2([0,\infty[^2)$. That is a necessary condition for a kernel, in order to make the induced integral operator *not* compact (if $k \in L^2([0,\infty[^2)$ then K was a Hilbert–Schmidt operator which are always compact, see Section 4.2).

An equivalent way to define K is

$$Kf = \sum_{n\in\mathbb{N}} (f, 1_{[n-1,n]})1_{[n-1,n]}.$$

Since $(1_{[n-1,n]})$ is an orthonormal basis for the range of K, we see from Theorem 2.17 that K is bounded but *not* compact.

APPENDIX C

On the Eigenvalues of Integral Operators

This chapter contains an excerpt from the article *"On the Characteristic Values of Linear Integral Equations"* by Hille and Tamarkin, 1931 (see [HT31]). This 76 pages long article contains a summary table which is presented in the following.

The notation has been changed a little in order to agree with the notation of this thesis. In the article, they let r_n be the absolute value of a characteristic value which corresponds to the reciprocal of an eigenvalue λ_n. Therefore, expressions like

$$r_n n^{-\sigma} \to \infty \quad \text{and} \quad r_n > R^{\frac{1-\epsilon}{4}n},$$

are replaced here by

$$|\lambda_n| = \mathcal{O}(n^{-\sigma}) \quad \text{and} \quad |\lambda_n| < R^{\frac{\epsilon-1}{4}n},$$

respectively.

Although they talk of general L^2 kernels, it seems like every kernel must be defined on a closed and bounded region, $[a,b] \times [a,b]$ where a and b are finite. The article makes no remarks on the *existence* of eigenvalues. So only when *assuming* there is an infinite sequence of eigenvalues does the results make sense.

The logarithms used in the table, $\log$, are natural logarithms. Furthermore, expressions like $\log_k$, $k = 1, 2, \ldots$, should be interpreted as

$$\log_2(x) = \log(\log(x)), \quad \log_3(x) = \log(\log(\log(x))), \quad \ldots.$$

In the last four rows in the table, no constant or $\mathcal{O}$-notation appears. This seems strange since a simple scaling of a given kernel will make it fit into the same class but the eigenvalues will be scaled accordingly.

The table will now be introduced. Let K, K_1 and K_2 be integral operators with kernels $k, k_1, k_2 \in L^2([a,b]^2)$ respectively. Unless stated otherwise, we use $a = 0$ and $b = 2\pi$. The sequence (λ_n) denotes the eigenvalues of K, repeated according to multiplicity and ordered such that $|\lambda_1| \geq |\lambda_2| \geq \cdots$.

The quantity ϵ denotes an arbitrarily small fixed positive quantity, not necessarily the same in all formulas.

The table refers at some places to certain kernel classes. These will be defined in the following.

The class $\Upsilon_a(\beta, q)$

Let (ϕ_n) be an orthonormal basis for $L^2(I)$ and set $r_n(t) = (k(\cdot,t), \phi_n)$, $n \in \mathbb{N}$.

For $\beta > 0$ and $q \geq 2$ we have that a kernel $k \in \Upsilon_a(\beta, q)$ if there exists an $n_0 \in \mathbb{N}$ such that the series

$$\sum_{n=n_0}^{\infty} n^{\beta} |r_n(t)|^q = \Omega(t),$$

converges for almost all $t \in J$ and its sum $\Omega(t)$ is integrable.

The class $\Upsilon_b(v, \alpha, p_1, p_2)$

We must have $v \geq 0$ is integer and that $\alpha, p_1, p_2 \in \mathbb{R}$ with $0 < \alpha \leq 1$ and $p_1, p_2 > 1$. If $k \in \Upsilon_b(v, \alpha, p_1, p_2)$ then the partial derivatives

$$\partial_s k(s,t), \ldots, \partial_s^v k(s,t)$$

must exist for almost all t. In case $v > 0$ the functions

$$\partial_s^n k(s,t), n = 0, 1, \ldots, v-1$$

must be continuous in s on $0 \leq s \leq 2\pi$ for almost all t, and $\partial_s^v k(s,t)$ must be representable in the form

$$\partial_s^v k(s,t) = \begin{cases} \frac{1}{2\pi\Gamma(\alpha)} \int_0^{2\pi} g(z,t)\Psi_\alpha(s-z)dz, & \alpha < 1, \\ \int_0^s g(z,t)dz + C(t), & \alpha = 1, \end{cases}$$

where G fulfills that

$$\int_0^{2\pi} \left(\int_0^{2\pi} |g(s,t)|^{p_1} \right)^{p_2} dt$$

exists. The function Ψ_α has the form

$$\Psi_\alpha(x) = \begin{cases} 2\pi \lim_{k\to\infty} \left[\sum_{n=0}^{\infty} (x + 2\pi n)^{\alpha-1} - \frac{1}{\alpha}(2\pi)^{\alpha-1} k^{\alpha} \right], & 0 < \alpha < 1, \\ \pi - x, & \alpha = 1, \end{cases}$$

and is 2π-periodic, $\Psi_\alpha(x + 2\pi) = \Psi_\alpha(x)$.

The class $\mathrm{Lip}\,(v, \alpha, p, q)$

We must have $v \geq 0$ integer, $0 < \alpha < 1$ and $1 < p \leq 2 \leq q$. A kernel $k \in \mathrm{Lip}\,(v, \alpha, p, q)$ if, for almost all t, the partial derivatives

$$\partial_s^n k(s,t), \quad n = 1, 2, \ldots, v$$

exist and, in case $v \geq 1$, the functions

$$\partial_s^n k(s,t), \quad n = 1, 2, \ldots, v-1$$

are continuous in s for fixed t. Furthermore, the derivative $\partial_s^v k(s,t)$, considered as a periodic function of s outside the interval $(0, 2\pi)$ satisfies the condition

$$\int_0^{2\pi} |\partial_s^v k(s+\epsilon, t) - \partial_s^v k(s,t)|^p \, ds < g(t)\epsilon^{\alpha p},$$

where $g \in L^q$ and $\epsilon \geq 0$ is sufficiently small.

The class $\mathrm{Lip}_1(v,p)$

If $k \in \mathrm{Lip}_1(v,p)$ then $k \in \mathrm{Lip}\left(v, 0, p, \frac{1}{p-1}\right)$ with $v > 0$ and $1 < p \leq 2$.

The class $\mathrm{Lip}_2(v,p)$

If $k \in \mathrm{Lip}_2(v,p)$ then $k \in \mathrm{Lip}\left(v, 1, p, \frac{1}{p-1}\right)$ with $v \geq 0$ and $1 < p \leq 2$.

The class $\mathrm{Lip}_3(v)$

If $k \in \mathrm{Lip}_3(v)$ then $k \in \mathrm{Lip}\,(v, 0, 1, \infty)$ with $v > 0$.

The class $\mathrm{Lip}_4(v)$

If $k \in \mathrm{Lip}_4(v)$ then $k \in \mathrm{Lip}\,(v, 1, 1, \infty)$ with $v \geq 0$.

The class $\Upsilon_c(v, l, \alpha)$

Let v and l be non-negative integers and $0 \leq \alpha \leq 1$. A kernel $k \in \Upsilon_c(v, l, \alpha)$ if, for almost all t, the partial derivatives

$$\partial_s^n k(s,t), \quad n = 1, 2, \ldots, v$$

exist and, in case $v > 0$, the derivatives

$$\partial_s^n k(s,t), \quad n = 1, 2, \ldots, v-1$$

are continuous in s for t fixed. The derivative

$$\partial_s^v k(s,t) = G(s,t)$$

must satisfy $G \in L^2$ and when setting

$$\begin{aligned} g_0(s,t,z) &= G(s+2z,t) + G(s-2z,t) - 2G(s,t), \\ g_i(s,t,z) &= \int_0^z g_{i-1}(s,t,w)dw, i = 1, 2, \ldots, l, \\ G_i(s,t,z) &= i! z^{-i} g_i(s,t,z), \end{aligned}$$

we have for almost all s, t,

$$\int_0^\tau |G_l(s,t,z)|dz < \gamma_\tau(s,t)\tau^{1+\alpha}, \quad 0 \leq \alpha \leq 1,$$

where $\gamma_\tau \in L^2$ for $\tau > 0$ and $\|\gamma_\tau\|$ is bounded as $\tau \to 0$.

The class $\Upsilon_d(R)$

The kernel $k \in \Upsilon_d(R)$ if $k(s,t)$ is analytic in s for almost all t in the interior of an ellipse in the complex s-plane, whose foci are at the points ± 1 and whose sum of semi-axis is R, and that for all such values of s we have

$$|k(s,t)| \leq M(t), \quad M \in L^2.$$

The table now follows. For shorter expressions, we introduce the auxiliary quantity

$$\sigma = v + \alpha + 1 - \frac{1}{p}.$$

	Classification	Behaviour of eigenvalues
1	$k \in L^2$	$\sum_{n=1}^{\infty} \lvert\lambda_n\rvert^2 < \infty$
2	$K = K_1 K_2$	$\sum_{n=1}^{\infty} \lvert\lambda_n\rvert < \infty$
3	$k \in L^1 \cap L^2$ is Hermitian semi-definite	$\sum_{n=1}^{\infty} \lvert\lambda_n\rvert < \infty$
4	$k \in \Upsilon_a(\beta, q)$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-\frac{\beta+1}{q}}\right)$
5	$k \in \Upsilon_b\left(v, \alpha, p, \frac{1}{p-1}\right)$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-\sigma}\right)$
6	$k \in \mathrm{Lip}\left(v, \alpha, p, \frac{1}{p-1}\right)$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-\sigma}(\log n)^{v+\alpha}\right)$
7	$k \in \mathrm{Lip}_1(v, p)$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-(v+1-\frac{1}{p})}\right)$
8	$k \in \mathrm{Lip}_2(v, p)$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-(v+2-\frac{1}{p})}\right)$
9	$k \in \mathrm{Lip}_3(v)$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-v}(\log n)^{v+\frac{1}{2}}\right)$
10	$k \in \mathrm{Lip}_4(v)$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-v-1}(\log n)^{v+\frac{1}{2}}\right)$
11	$k \in \Upsilon_c(v, l, \alpha)$, $0 \le \alpha < 1$ $\alpha = 1$	$\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-v-\alpha-\frac{1}{2}}(\log n)^{v+\alpha}\right)$ $\lvert\lambda_n\rvert = \mathcal{O}\left(n^{-v-\frac{3}{2}}(\log n)^{v+2}\right)$
12	$k \in \Upsilon_d(R)$	$\lvert\lambda_n\rvert < R^{\frac{\epsilon-1}{4}n}$
13	$k(s,t) = \sum_{n=1}^{\infty} s^n \kappa_n(t), \gamma \in L^2$, $\lvert\kappa_n(t)\rvert < \gamma(t) e^{-\left(\frac{n}{\tau+\epsilon} \log_k n\right)}$	$\lvert\lambda_n\rvert < e^{\left(\frac{\epsilon-1}{4\tau} n \log_k n\right)}$, $k = 1, 2, \ldots$
14	$k(s,t) = \sum_{n=1}^{\infty} s^n \kappa_n(t), \gamma \in L^2$, $\lvert\kappa_n(t)\rvert < \gamma(t) \exp\left[-n^{1+\frac{1}{\tau+\epsilon}}\right]$	$\lvert\lambda_n\rvert < \exp\left[-\tau_0 n^{\left(\frac{\tau+1}{\tau}-\epsilon\right)}\right]$, $\tau_0 = \tau(\tau+1)^{\frac{\tau+1}{\tau}}(2\tau+1)^{-\frac{2\tau+1}{\tau}}$
15	$k(s,t) = \sum_{n=1}^{\infty} s^n \kappa_n(t), \gamma \in L^2$, $\lvert\kappa_n(t)\rvert < \gamma(t) \exp\left[-n e_{k-1}^{\left(n^{\frac{1}{\tau+\epsilon}}\right)}\right]$	$\lvert\lambda_n\rvert < e_k^{\left(-n^{\frac{1}{\tau}-\epsilon}\right)}$, $k = 2, 3, \ldots$

APPENDIX D

Examples

This chapter will exemplify different aspects of the theory presented in this thesis.

D.1 A simple Volterra Operator — Integration

Possibly one of the simplest non self-adjoint operators is the following Volterra operator:

$$(Kf)(s) = \int_0^s f(t)dt, \quad 0 \leq s \leq 1.$$

The kernel is thus $k(s,t) = 1$ for $0 \leq t \leq s \leq 1$ and 0 otherwise. The adjoint operator is easily obtained by using the conjugate transposed kernel $k^*(s,t) = \overline{k(t,s)}$:

$$(K^*f)(s) = \int_0^1 k^*(s,t)f(t)dt = \int_s^1 f(t)dt, \quad 0 \leq s \leq 1.$$

The Singular Value Expansion

To find the singular value expansion, we must calculate K^*K:

$$(K^*Kf)(s) = \int_s^1 \int_0^z f(t)dtdz.$$

The singular values are now obtained by solving the eigenvalue problem $K^*Kv = \mu^2 v$, which is equivalent to solving the ordinary differential equation

$$\mu^2 v'' + v = 0$$

with boundary conditions $v(1) = v'(0) = 0$ (see [Kre99, p. 280]). The nontrivial solutions are given by

$$\mu_n = \frac{2}{(2n-1)\pi}, \quad v_n(t) = \sqrt{2}\cos\left(\frac{(2n-1)\pi t}{2}\right), \quad n \in \mathbb{N}.$$

The singular value expansion is completed by calculating

$$u_n(s) = \frac{1}{\mu_n}(Kv_n)(s) = \sqrt{2}\sin\left(\frac{(2n-1)\pi s}{2}\right).$$

Discretization

Assume we wish to discretize this operator using box functions, leading to an $\mathbf{A} \in \mathbb{R}^{N\times N}$ matrix. As seen in Section A.1, this can be done in the following way:

$$\mathbf{A}_{i,j} = N \int_{\frac{i-1}{N}}^{\frac{i}{N}} \int_{\frac{j-1}{N}}^{\frac{j}{N}} k(s,t)dtds.$$

When $j > i$ the kernel will be zero in the integration interval. If $j < i$ the kernel will be constantly 1, obtaining $\mathbf{A}_{ij} = \frac{1}{N}$. When $i = j$ the integration square is divided along the diagonal, one part with $k(s,t) = 1$ and one with $k(s,t) = 0$. This leads to $\mathbf{A}_{ij} = \frac{1}{2N}$.

All in all we arrive at a matrix looking like this:

$$\mathbf{A} = \frac{1}{2N}\begin{bmatrix} 1 & & & \\ 2 & 1 & & \\ 2 & 2 & 1 & \\ \vdots & \vdots & \ddots & \ddots \end{bmatrix} \in \mathbb{R}^{N\times N}.$$

Notice that all N eigenvalues of this matrix are $\lambda = \frac{1}{2N}$. So for increasing N every eigenvalue of $\mathbf{A}$ converges to zero. This is what could be expected since K is a Volterra operator that has no eigenvalues at all ($\lambda = 0$ is not an eigenvalue since K only has the trivial null-space).

Let us instead try to discretize with respect to a basis consisting of the singular vectors (v_n) and (u_n):

$$\mathbf{B}_{i,j} = (Kv_j, u_i) = (\mu_j u_j, u_i) = \mu_j \delta_{ij} = \mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_N)$$
$$\text{where } \mu_i = \frac{2}{(2n-1)\pi}.$$

So the non self-adjoint compact operator has lead to a simple diagonal matrix! The eigenvalues here are clearly just $\lambda_i = \mu_i$.

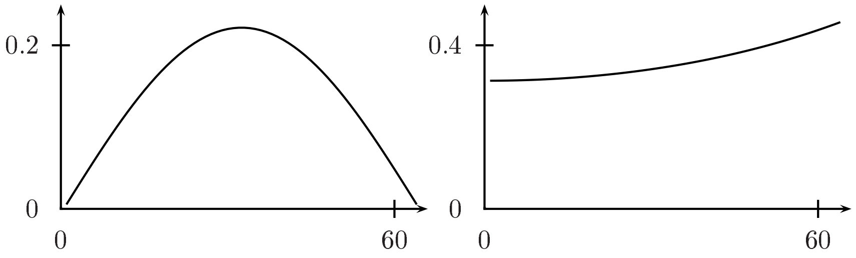


Figure D.1: Plots of the solution $\mathbf{x}$ *(left) and the right-hand side* $\mathbf{b}$ *(right) from the* baart *test problem (both are discretized versions).*

D.2 baart — convergence plots

The test problem used here is called baart in the REGULARIZATION TOOLBOX, see [Han98b]. The integration kernel is

$$k(s,t) = e^{s\cos t}, \quad (s,t) \in [0, \tfrac{\pi}{2}] \times [0,\pi],$$

where a solution with corresponding right-hand side are given by

$$f(t) = \sin t \quad \text{and} \quad g(s) = 2\frac{\sin s}{s},$$

respectively. Discretized versions of these can be seen in Figure D.1. The discretization used was Galerkin discretization with box basis functions.

Eigenvalues

Important to note is the fact that the domain and range are *different* for $K : L^2([0,\pi]) \to L^2([0,\frac{\pi}{2}])$. It therefore makes no sense to talk about eigenvalues of K. But when discretizing using N basis functions in each space, we arrive at a square matrix $\mathbf{A}$ for which it makes perfect sense to talk of eigenvalues. Let now $(\phi_n)_{n=1}^N$ and $(\psi_n)_{n=1}^N$ be sequences of orthonormal box functions in $L^2([0,\pi])$ and $L^2([0,\frac{\pi}{2}])$ respectively. Note that $\psi_n(t) = \phi_n(2t)$ for $0 \le t \le \frac{\pi}{2}$. Assume now that

$\mathbf{A}$ has an eigenvalue λ with corresponding eigenvector $\boldsymbol{\varphi}$. This means that

$$\sum_{j=1}^{N} \mathbf{A}_{i,j}\boldsymbol{\varphi}_j = \lambda\boldsymbol{\varphi}_i, \quad \text{for } i = 1,2,\dots,N.$$

Let now $\widetilde{K}_N$ be the approximation to K given by $\mathbf{A}$ and let $\varphi = \sum_{j=1}^N \boldsymbol{\varphi}_j\phi_j$ (see Section 6.1). We now get

$$\begin{aligned}\widetilde{K}_N\varphi(s) &= \sum_{i=1}^{N}\sum_{j=1}^{N}(K\phi_j,\psi_i)(\varphi,\phi_j)\psi_i(s) = \sum_{i=1}^{N}\sum_{j=1}^{N}\mathbf{A}_{i,j}\boldsymbol{\varphi}_j\psi_i(s)\\ &= \lambda\sum_{i=1}^{N}\boldsymbol{\varphi}_i\psi_i(s) = \lambda\sum_{i=1}^{N}(\varphi,\phi_i)\phi_i(2s) = \lambda\varphi(2s)\end{aligned} \tag{D.1}$$

So the operator $\widetilde{K}_N$ does not, of course, have eigenvectors, but a similar property that also *scales* the output vector. In general, we have

$$\int_J k(s,t)\varphi(t)dt = \lambda\varphi(2s) \quad \Leftrightarrow \quad \int_J k(\tfrac{1}{2}s,t)\varphi(t)dt = \lambda\varphi(s).$$

This means that the eigenvector from Equation (D.1) actually is an approximate eigenvector to the integral operator with kernel

$$k(s,t) = e^{\frac{1}{2}s\cos t}, \quad (s,t) \in [0,\tfrac{\pi}{2}]^2 \tag{D.2}$$

instead. Notice that a similar "trick" can be used whenever the basis functions used are identical except for translation and/or scaling.

Running GMRES

We now wish to run GMRES on the discretized version of the problem introduced in the beginning. Before proceeding, we compute some important quantities relating to this problem.

Figure D.2 shows the singular values and the eigenvalues of the coefficient matrix, $\mathbf{A} \in \mathbb{R}^{64\times 64}$. Notice how both decay (approximately)

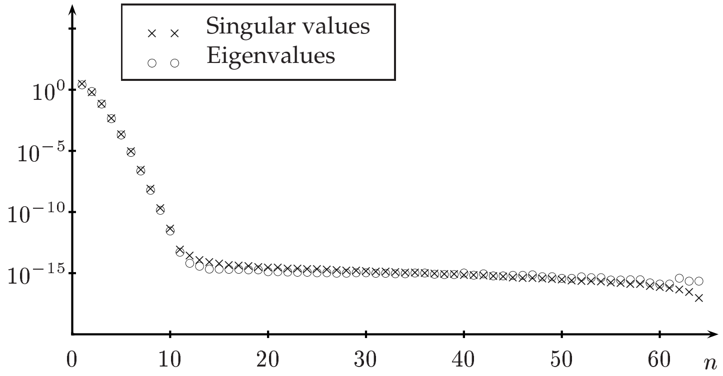


Figure D.2: *Singular values and eigenvalues of the* baart *test problem.*

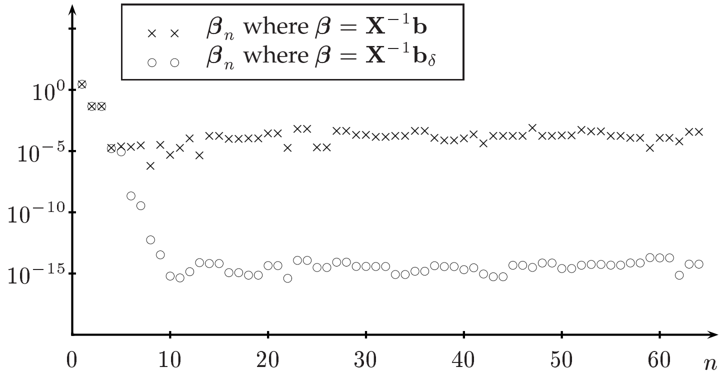


Figure D.3: *The eigenvector components of the right-hand sides* $\mathbf{b}$ *and* $\mathbf{b}_\delta$. $\mathbf{X}$ *denotes a matrix such that* $\mathbf{X}^{-1}\mathbf{A}\mathbf{X} = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$ *where* $|\lambda_1| \geq |\lambda_2| > \cdots$.

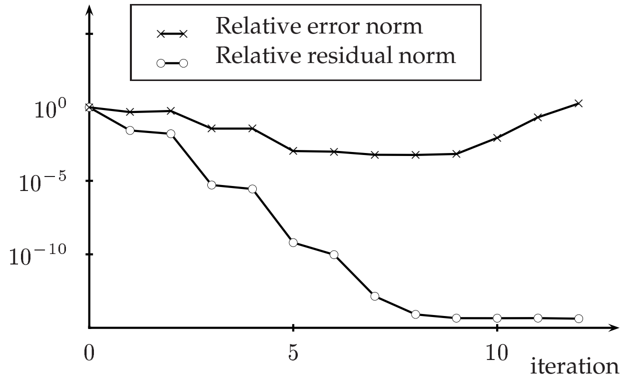


Figure D.4: *Convergence of* baart *using GMRES. The error norm increases from around iteration 8.*

exponentially until they hit a level around 10^{-15} due to noise in the coefficient matrix and floating point rounding errors. Figure D.3 shows the eigenvector components of the right-hand side $\mathbf{b}$.

Figure D.4 shows the result of running GMRES on this problem. For each iteration, the relative residual norm and relative error norm are shown. Note that the error is measured against the true solution shown earlier (which is extraordinarily known). From the figure it can also be seen that the decay rate of the residual norm is approximately exponential. Important to note is also that the error norm does not decay as fast as the residual norm and that it begins to grow at some point. A good rule of thumb seems to be that the number corresponding to when the right-hand side eigenvector components hit machine precision (or a larger error level) dictates the maximal number of iterations that should be used.

A convergence plot is also shown in Figure D.5. The problem is the same as above except that the right-hand side used is

$$\mathbf{b}_\delta = \mathbf{b} + 10^{-5}\|\mathbf{b}\|_2\mathbf{e},$$

where $\mathbf{e}$ is a vector of normally distributed noise with mean value 0

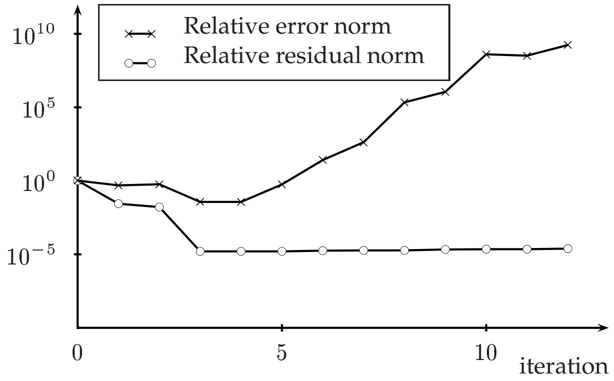


Figure D.5: *Convergence of* **baart** *using GMRES. Notice the semiconvergence: The error norm (from the* true *solution) increases from around iteration 4.*

and standard deviation 1. The right-hand side eigenvector components are also seen in Figure D.3.

In the convergence plot we have again that the error norm is measured against the true solution $\mathbf{x}$ and not the solution corresponding to $\mathbf{b}_\delta$. A clear example of semiconvergence can be seen. From around iteration 4 the solution vector starts to approach $\mathbf{A}^{-1}\mathbf{b}_\delta$ together with other unwanted inaccuracies. Note that this corresponds to where the right-hand side eigenvector components began to level out.

Theory and Practise

We will now look at the asymptotic behavior of the eigenvalues of the integral operator induced by the kernel in Equation (D.2). Using the well-known Taylor expansion of e^x, we get

$$k(s,t) = e^{\frac{1}{2}s\cos t} = \sum_{n=1}^{\infty} \frac{1}{n!}\left(\tfrac{1}{2}s\cos t\right)^n = \sum_{n=1}^{\infty} s^n \kappa_n(t),$$
$$\text{where } \kappa_n(t) = \tfrac{1}{n!}\tfrac{1}{2^n}\cos^n t.$$

Aiming to use the bound in row 13 of the table on page 145, we recall the Stirling formula for the factorial function:

$$n! > \sqrt{2\pi n}\, n^n e^{-n}.$$

Although these quantities are very close, the inequality actually holds (see [Knu97, page 115]). We now get

$$|\kappa_n(t)| < \frac{1}{\sqrt{2\pi n}\, n^n e^{-n}} \frac{1}{2^n} = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\ln(n) - n\ln(n) + n - \ln(2)n}$$
$$< \frac{1}{\sqrt{2\pi}} e^{-n(\ln(n) - 1 + \ln(2))} \lesssim \gamma e^{-n\ln(n)}$$

for a constant γ and from some value of n. According to the table, the eigenvalues should now be bounded by

$$|\lambda_n| < e^{-\frac{1}{4}n\ln(n)}.$$

The computed eigenvalues and this bound are plotted in Figure D.6. The bound is seen to be quite pessimistic, although the shape of the curve seems correct. Also in the figure is plotted

$$e^{-n\ln(n)},$$

which is seen to fit *very* well. So for this example, the factor in the exponent was too pessimistic. Note that the computed eigenvalues level out due to lack of floating point precision.

We will now apply the residual norm bounds derived in Section 7.4.1 to the convergence curve seen in Figure D.4. By computing the eigenvalues and $\boldsymbol{\beta}$ ($= \mathbf{X}^{-1}\mathbf{b}$, the eigenvector components of the right-hand side), we can explicitly calculate the polynomial $p_k^{\max}$ (see Equation (7.15)) and compute the bound for each k. This lead to one of the bounds seen in Figure D.7. Also shown in the figure is the approximate bound computed using only $\boldsymbol{\beta}$ (since $p_k = 1$), see Equation (7.16) page 102. This bound is seen to be almost identical to the previous. Apart from a constant factor, both bounds are seen to follow the shape of the convergence curve very tightly.

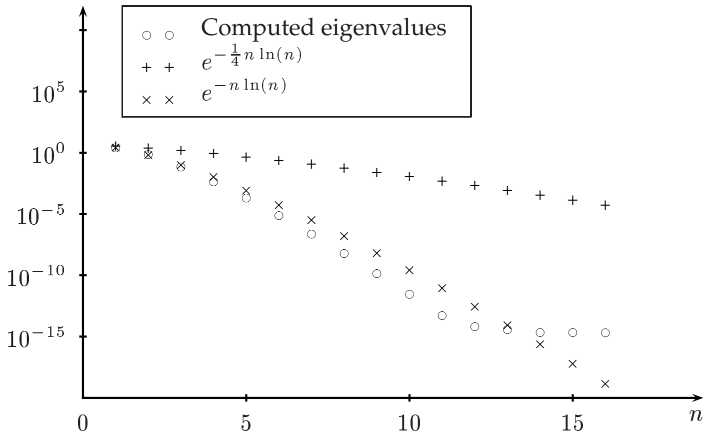


Figure D.6: *Expected behavior of the eigenvalues for the* baart *test problem.*

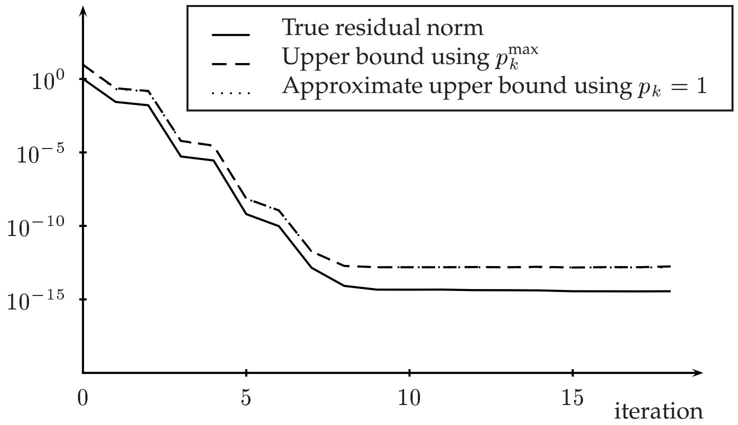


Figure D.7: *Expected convergence for the* baart *test problem, no noise.*

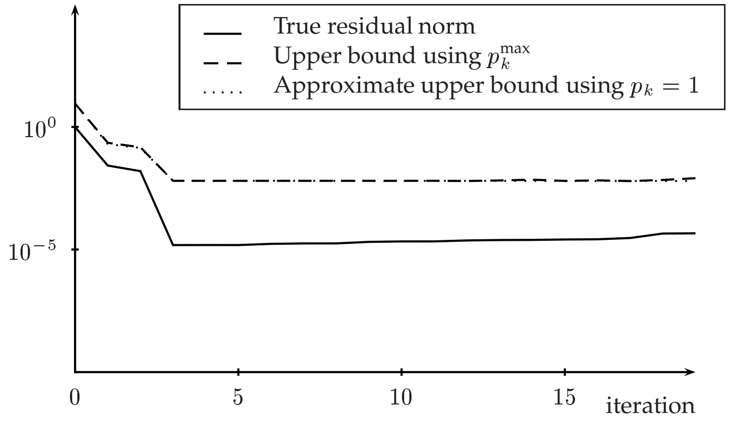


Figure D.8: *Expected convergence for the* baart *test problem, with noise.*

The results of trying out these bounds on the problem influenced by noise can be seen in Figure D.8. Here the bounds are still good, but especially in the beginning. As soon as the $\boldsymbol{\beta}_i$'s do not decay fast enough, the bounds become more pessimistic.

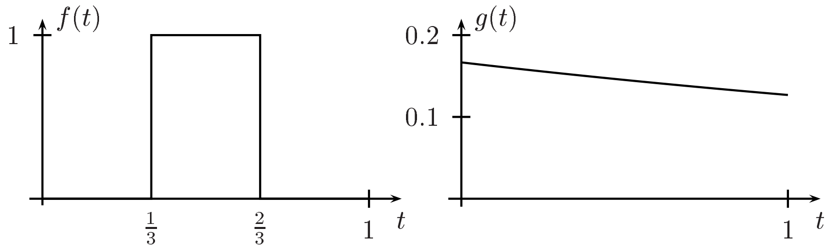


Figure D.9: *Plots of f and g from the* wing *test problem.*

D.3 wing — when GMRES fails

This test problem is also from the REGULARIZATION TOOLBOX. The integral kernel has the form

$$k(s,t) = t\, e^{-st^2}, \quad (s,t) \in [0,1]^2.$$

A solution and right-hand side are given by

$$f(t) = \begin{cases} 1, & \text{if } t_1 \le t \le t_2 \\ 0, & \text{otherwise} \end{cases} \quad \text{and} \quad g(s) = \frac{e^{-st_1^2} - e^{-st_2^2}}{2s},$$

respectively. For this example we use $t_1 = \frac{1}{3}$ and $t_2 = \frac{2}{3}$ and the goal here is clearly to attempt restoring the discontinuities of f, given the right-hand side g. A plot of f and g can be seen in Figure D.9.

Discretization

A 64×64 coefficient matrix was computed using box basis functions.[1] Figure D.10 shows the eigenvalues and singular values of the coefficient matrix. Note that these values level off at about 10^{-8}, somewhat higher than the machine precision. This is probably due to noise in the coefficient matrix, caused by the numerical discretization.

[1]See Chapter F in the appendix for information on the software used for numerical integration.

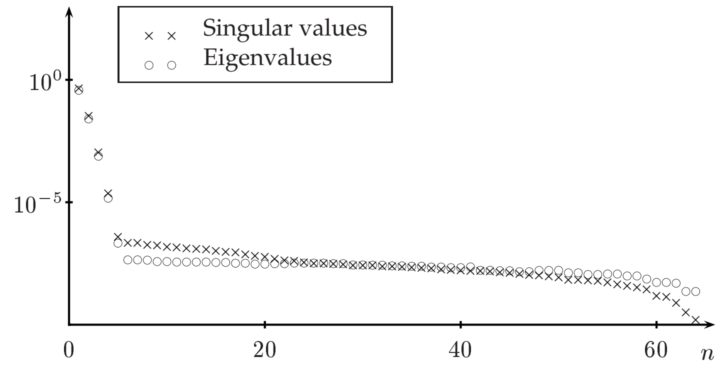


Figure D.10: *Singular values and eigenvalues of the* wing *test problem.*

Running GMRES

GMRES was run on this problem. Figure D.11 shows the first 5 approximate solution vectors. They all fail to come close to the true solution and the fifth vector is seen to be influenced by noise. This means that there is very little hope that a better solution will be found.

Also in the figure are 5 eigenvectors corresponding to the 5 largest eigenvalues. Although GMRES assembles its solutions from Krylov subspaces, we know from the convergence analysis of GMRES (see Section 7.4.1) that the approximate solutions are built (mostly) from the eigenvectors corresponding to the largest eigenvalues. This is clearly seen from the (visual) resemblance between the approximate solutions and the eigenvectors.

The poor convergence is also made clear in Figure D.12. Here, the residual norm and error norm are shown for each iteration. The relative error norm actually never manages to come below 1! This is clearly because the vectors of the Krylov subspaces, from which the approximate solutions are built, are totally unfit to assemble the discontinuous solution.

Figure D.13 shows a similar convergence plot, but here the dis-

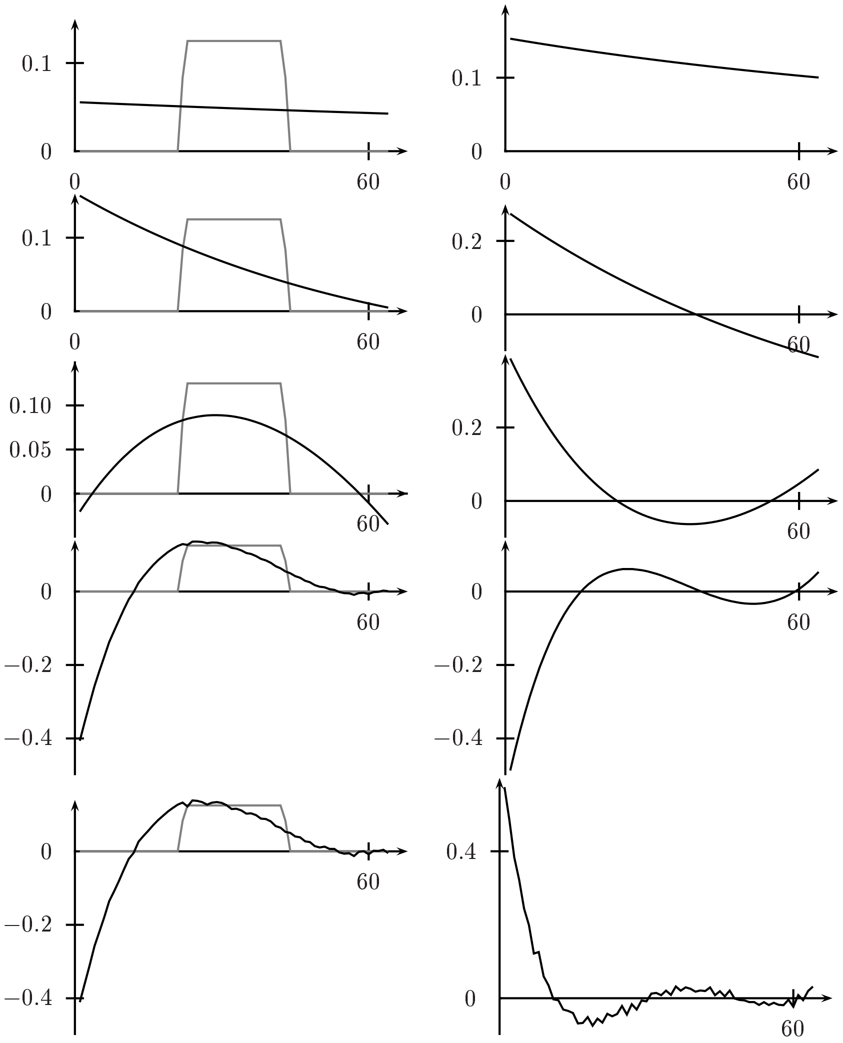


Figure D.11: *The left column shows iteration vectors* $\mathbf{x}^{(1)}$ *to* $\mathbf{x}^{(5)}$*, the true solution is shown in gray. The column on the right shows 5 eigenvectors corresponding to the 5 largest eigenvalues.*

cretization was computed using cosine basis function.[2] The problems are seen to be the same. The error norm is actually seen to increase instead of decrease.

[2]The coefficient matrix is only 16×16 since accurate numerical integration with high frequency cosine functions proved very difficult.

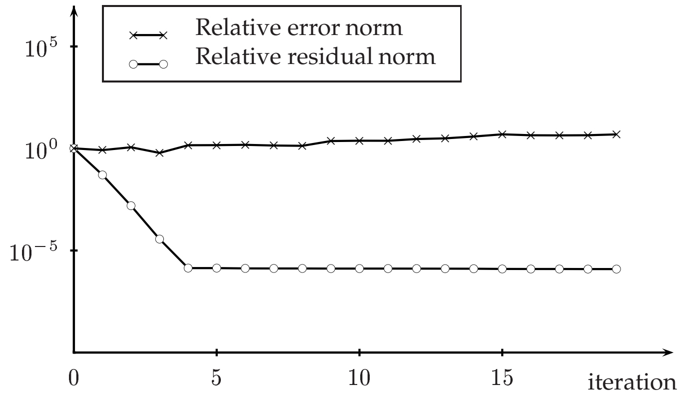


Figure D.12: *Convergence of* wing *using GMRES. Discretization was done using box basis functions.*

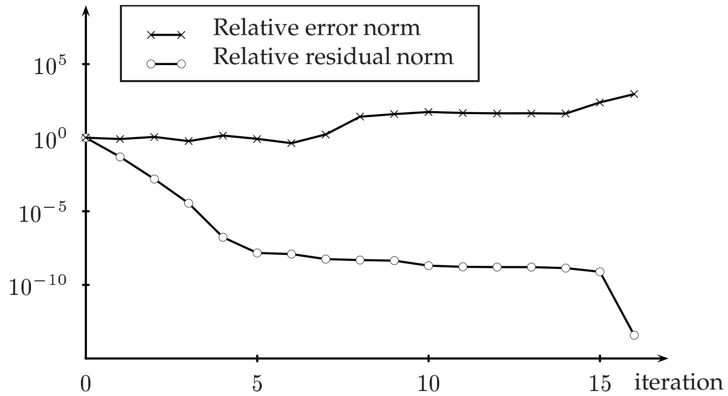


Figure D.13: *Convergence of* wing *using GMRES. Discretization was done using cosine basis functions.*

D.4 **deriv2** — Discontinuous derivative

Consider the self-adjoint operator $K : L^2([0,1]) \to L^2([0,1])$ with kernel

$$k(s,t) = \begin{cases} t(s-1), & t \le s \\ s(t-1), & t > s. \end{cases} \tag{D.3}$$

This is the so-called Green's function to the boundary value problem for the ordinary differential equation[3]

$$g''(s) = f(s), \quad 0 \le s \le 1,$$

with boundary condition $g(0) = g(1) = 0$. It can be shown that for each function $f \in C([0,1])$ a unique solution $g \in C^2([0,1])$ to this problem is given by $g = Kf$.

The eigenvalue problem $K\varphi = \lambda\varphi$ is equivalent to the differential equation

$$\lambda\varphi''(t) - \varphi(t) = 0, \quad 0 \le x \le 1,$$

with boundary condition $\varphi(0) = \varphi(1) = 0$. The nontrivial solutions to this problem are

$$\lambda_n = -\frac{1}{\pi^2 n^2}, \qquad \varphi_n(t) = \sqrt{2}\sin(n\pi t).$$

We will now try, from the kernel (D.3) in alone, to predict the asymptotic decay of the eigenvalues. The kernel itself is obviously continuous and the first partial derivative with respect to s is

$$\partial_s k(s,t) = \begin{cases} t, & t \le s \\ t-1, & t > s. \end{cases}$$

Since obviously $\|\partial_s k(\cdot,t)\|_1 < \infty$ for all t, all the requirements of Theorem 4.10 are fulfilled and we have

$$|\lambda_n| = \mathcal{O}\left(n^{-1}(\log n)^{\frac{3}{2}}\right). \tag{D.4}$$

[3] A closely related example is discussed in [Kre99] page 275.

But that is not a very tight bound. So let us try with the class $\mathrm{Lip}_1(1,2)$ instead. Since $p=2$ we see that the condition

$$\int_0^1 \left|\partial_s k(s+\epsilon,t) - \partial_s k(s,t)\right|^2 ds < g(t) \tag{D.5}$$

must be satisfied with $g \in L^1([0,1])$ and $\epsilon > 0$ sufficiently small. Because $\partial_s k(s,t)$ shall be considered as a periodic function of s outside the interval $[0,1]$, we have

$$\partial_s k(s+\epsilon,t) - \partial_s k(s,t) = \begin{cases} 0, & \text{for } s < t-\epsilon \\ 1, & \text{for } t-\epsilon \le s < t \\ 0, & \text{for } t \le s \end{cases}$$

which leads to

$$\int_0^1 \left|\partial_s k(s+\epsilon,t) - \partial_s k(s,t)\right|^2 ds = \int_{t-\epsilon}^t 1 ds = \epsilon,$$

for every $\epsilon > 0$.

So the requirement in Equation (D.5) is satisfied with g a constant function. This provides us with the bound

$$|\lambda_n| = \mathcal{O}\left(n^{-(1+1-\frac{1}{2})}\right) = \mathcal{O}\left(n^{-\frac{3}{2}}\right).$$

This bound is better than the previous in (D.4), but it is not the tightest possible.

Let us turn to the class $\mathrm{Lip}_2(1,2)$. Our kernel is doomed to not satisfy the conditions because it would lead to a $\mathcal{O}(n^{-\frac{5}{2}})$ bound, but let us see what goes wrong. The condition here becomes

$$\int_0^1 \left|\frac{\partial_s k(s+\epsilon,t) - \partial_s k(s,t)}{\epsilon}\right|^2 ds < g(t), \tag{D.6}$$

for some $g \in L^1([0,1])$. Evaluating the left-hand side we get

$$\int_0^1 \left|\frac{\partial_s k(s+\epsilon,t) - \partial_s k(s,t)}{\epsilon}\right|^2 ds = \int_{t-\epsilon}^t \left|\frac{1}{\epsilon}\right|^2 ds = \frac{1}{\epsilon},$$

for every $\epsilon > 0$. This clearly shows that no function $g \in L^1$ exists such that $1/\epsilon < g(t)$ for all $t \in [0,1]$ and all ϵ. So, as expected, the kernel k does not lie in the $\mathrm{Lip}_2(1,2)$ class.

Before giving up trying to find a tighter bound, let us try with the class $\Upsilon_b(v,1,p,\frac{1}{1-p})$. The requirements for this class are expressed in Theorem 4.9, saying that we must find functions g and C that fulfill

$$\partial_s k(s,t) = \int_0^s g(z,t)dz + C(t). \tag{D.7}$$

Focusing on the region $t < s$ we now wish to differentiate the above expression on both sides with respect to s. If $G(s,t)$ is a function that fulfills $\frac{\partial}{\partial s}G(s,t) = g(s,t)$ then we get

$$0 = \frac{\partial}{\partial s}\left(G(s,t) - G(0,t) + C(t)\right) = g(s,t),$$

for $t < s$. Considering the region $t > s$ we see that it leads to an analogous conclusion, $g(s,t) = 0$ for $t > s$. This also makes sense, since $\partial_s k(s,t)$ is constant in the s-direction on each triangular region. Since the added function $C(t)$ can not represent the jump in the s-direction along the diagonal, it must be handled by $g(s,t)$. But this requires using distribution theory and the Dirac delta-function,

$$g(s,t) = -\delta(t-s),$$

such that

$$\int_0^s g(z,t)dt = -\int_0^s \delta(t-z)dz = \begin{cases} 0, & \text{for } t \le s \\ -1, & \text{for } t > s \end{cases}.$$

This would, with $C(t) = t$, fulfill Equation (D.7). But g, being such a delta-function, does *not* fulfill the requirement that for some $1 < p \le 2$ we must have

$$\int_0^1 \left(\int_0^1 |g(s,t)|^p\right)^{\frac{1}{p-1}} dt < \infty$$

So in conclusion: The kernel k does *not* lie in the class Υ_b, either.

D.5 Image Deblurring — 2D Domain and Range

This example models atmospheric turbulence blurring of images. Given an (original) image, $f \in L^2([0,1]^2)$, the following expression computes the blurred image, $g \in L^2([0,1]^2)$:

$$g(x,y) = \int_0^1 \int_0^1 k(x-x', y-y') f(x',y') dx' dy', \quad (x,y) \in [0,1]^2, \quad \text{(D.8)}$$

where

$$k(x,y) = \frac{1}{\sqrt{2\pi}\sigma} e^{\left(-\frac{x^2+y^2}{2\sigma^2}\right)} = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x}{\sigma}\right)^2} e^{-\frac{1}{2}\left(\frac{y}{\sigma}\right)^2} = C k_x(x) k_y(y)$$

is a Gaussian point-spread function.

Note how both the domain and the range are two-dimensional, leading to a double integral.

Discretizing

Discretizing is not much different from the one-dimensional problems, though. We want to use box functions as basis functions, leading to $N \times N$ discrete images. The matrix entries (or more correctly, tensor entries) become, cf. Equation (A.1):

$$\begin{aligned} \mathbf{A'}_{i_x,i_y,j_x,j_y} &= \left(K(e_{j_x} \otimes e_{j_y}), e_{i_x} \otimes e_{i_y}\right) \\ &= \int_0^1 \int_0^1 \int_0^1 \int_0^1 k(x-x', y-y') e_{j_x}(x) e_{j_y}(y) e_{i_x}(x') e_{i_y}(y') dx' dy' dx dy \\ &= CN^2 \int_{\frac{j_x-1}{N}}^{\frac{j_x}{N}} \int_{\frac{i_x-1}{N}}^{\frac{i_x}{N}} k_x(x-x') dx' dx \int_{\frac{j_y-1}{N}}^{\frac{j_y}{N}} \int_{\frac{i_y-1}{N}}^{\frac{i_y}{N}} k_y(y-y') dy' dy. \end{aligned} \quad \text{(D.9)}$$

Since we thus have to compute integrals of the form $\int_a^b e^{-x^2} dx$, which are hard to handle analytically, we assume that N is large enough

to use the approximation

$$\int_a^b h(x) dx \simeq (b-a) h\left(\frac{a+b}{2}\right).$$

This leads to the following expression

$$\mathbf{A'}_{i_x,i_y,j_x,j_y} \simeq \frac{C}{N^2} k_x\left(\frac{j_x - i_x}{N}\right) k_y\left(\frac{j_y - i_y}{N}\right) = \frac{C}{N^2} \mathbf{B}_{i_x,j_x} \mathbf{B}_{i_y,j_y} \quad \text{(D.10)}$$

since $k_x = k_y$ and where

$$\mathbf{B}_{i,j} = e^{-\frac{1}{2}\left(\frac{j-i}{\sigma N}\right)^2}.$$

Since $\mathbf{B}$ has constant entries along the main- and each off-diagonal, it is called a *Toeplitz* matrix. Since it is furthermore symmetric, it is called a *Hankel* matrix.

We can now formulate Equation (D.8) in our discrete notation. Assume the original and blurred image are represented as

$$\tilde{f} = \sum_{i_y=1}^{N} \sum_{i_x=1}^{N} \mathbf{F}_{i_y,i_x} e_{i_x} \otimes e_{i_y}, \quad \text{and} \quad \tilde{g} = \sum_{j_y=1}^{N} \sum_{j_x=1}^{N} \mathbf{G}_{j_y,j_x} e_{j_x} \otimes e_{j_y},$$

respectively, then we have

$$\mathbf{G}_{j_y,j_x} = \sum_{i_y=1}^{N} \sum_{i_x=1}^{N} \mathbf{A'}_{i_x,i_y,j_x,j_y} \mathbf{F}_{i_y,i_x} \quad \text{(D.11)}$$

Because of the double sum, this cannot be expressed as a matrix equation. However, by "stacking" $\mathbf{F}$ and $\mathbf{G}$ into

$$\mathbf{x}_{(i_x-1)N+i_y} = \mathbf{F}_{i_y,i_x} \quad \text{and} \quad \mathbf{b}_{(j_x-1)N+j_y} = \mathbf{G}_{j_y,j_x},$$

and similarly for the $\mathbf{A}'$-tensor,

$$\mathbf{A}_{(i_x-1)N+i_y,(j_x-1)N+j_y} = \mathbf{A'}_{i_x,i_y,j_x,j_y}, \quad \text{(D.12)}$$

we get the well-known

$$\mathbf{b} = \mathbf{A}\mathbf{x}. \tag{D.13}$$

If Equation (D.12) is combined with (D.10), we get

$$\mathbf{A}_{(i_x-1)N+i_y,(j_x-1)N+j_y} = \frac{C}{N^2}\mathbf{B}_{i_x,j_x}\mathbf{B}_{i_y,j_y} \quad \Leftrightarrow$$

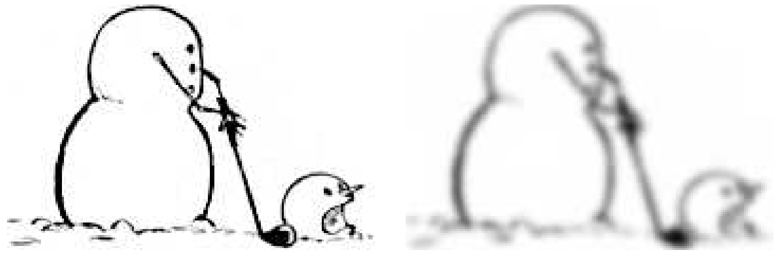
$$\mathbf{A} = \frac{C}{N^2}\begin{bmatrix} \mathbf{B}_{1,1}\mathbf{B} & \mathbf{B}_{1,2}\mathbf{B} & \cdots & \mathbf{B}_{1,N}\mathbf{B} \\ \mathbf{B}_{2,1}\mathbf{B} & \mathbf{B}_{2,2}\mathbf{B} & \cdots & \mathbf{B}_{2,N}\mathbf{B} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{B}_{N,1}\mathbf{B} & \mathbf{B}_{N,2}\mathbf{B} & \cdots & \mathbf{B}_{N,N}\mathbf{B} \end{bmatrix} = \frac{C}{N^2}\mathbf{B} \otimes \mathbf{B}$$

where the symbol $\otimes$, when connected with matrices, is called a *Kronecker product*.

Important to note

The transformation from Equation (D.11) into Equation (D.13) is always possible for "2D integral equations", i.e. when the domain and range are two-dimensional. The special structure of the kernel made it possible to separate the variables as in (D.9), which in turn made it possible to use the short Kronecker product notation. The special Toeplitz structure of the $\mathbf{B}$-matrices was due to the convolution type kernel.

The REGULARIZATION TOOLBOX also has a test problem, blur, that models this kind of image blurring. Using this, the Figure D.14 was created to illustrate image blurring.



(a) Original image (b) Blurred image

Figure D.14: *An example of image blurring (the image is from a Calvin and Hobbes comic strip).*

D.6 Degenerate Kernel — no Krylov solution

Consider the integral operator $K : L^2([0,\pi]) \to L^2([0,\pi])$ with the following degenerate kernel

$$\begin{aligned} k(s,t) &= \frac{4}{\pi}\cos(s)\cos(t) + \frac{4}{\pi}\cos(2s)\cos(2t) + \frac{2}{\pi}\cos(4s)\cos(3t) \\ &= \sum_{n=1}^{3} a_n(s) b_n(t), \end{aligned}$$

where

$$\begin{aligned} a_1(s) &= \frac{4}{\pi}\cos(s), & a_2(s) &= \frac{4}{\pi}\cos(2s), & a_3(s) &= \frac{2}{\pi}\cos(4s), \\ b_1(t) &= \cos(t), & b_2(t) &= \cos(2s), & b_3(t) &= \cos(3t). \end{aligned}$$

Eigenvalues and eigenfunctions

We now wish to find the eigenvalues and -functions of K by considering a matrix eigenvalue problem instead, see Section 4.3.3. An obvious basis that spans all of the above a- and b-functions is

$$e_n(t) = \sqrt{\frac{2}{\pi}}\cos(nt), \quad n = 1,2,3,4.$$

We now form a matrix $\mathbf{W} \in \mathbb{R}^{4\times 4}$ by calculating the entries $\mathbf{W}_{i,j} = (Ke_j, e_i)$:

$$\mathbf{W} = \begin{bmatrix} 2 & & & \\ & 2 & & \\ & & 0 & \\ & & 1 & 0 \end{bmatrix}.$$

This matrix has an eigenvalue $\lambda = 2$ with multiplicity 2 and index 1. The corresponding eigenspace is spanned by $[\ 1\ \ 0\ \ 0\ \ 0\ ]^T$ and $[\ 0\ \ 1\ \ 0\ \ 0\ ]^T$. It also has the eigenvalue $\lambda = 0$ with multiplicity 2 and index 2. The null-space is thus spanned by only one vector: $[\ 0\ \ 0\ \ 0\ \ 1\ ]^T$.

As shown in Section 4.3.3, the eigenvalues of the operator K are identical to those of $\mathbf{W}$. The eigenspace associated with $\lambda = 2$ is now spanned by e_1 and e_2 instead, and the eigenspace associated with $\lambda = 0$ is spanned by e_4. But since the operator K was considered defined on the infinite dimensional space $L^2([0,\pi])$ then we have $\lambda = 0$ is *also* an eigenvalue with multiplicity ∞. The corresponding eigenspace is clearly spanned by all vectors orthogonal to e_1, e_2, e_3, e_4.

Using GMRES

Since zero was an eigenvalue with index > 1 of the matrix $\mathbf{W}$ shown above, what happens if we run GMRES on the following system

$$\begin{bmatrix} 2 & & & \\ & 2 & & \\ & & 0 & \\ & & 1 & 0 \end{bmatrix} \mathbf{x} = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 1 \end{bmatrix} \tag{D.14}$$

with obvious solution $[\ 1\ \ 1\ \ 1\ \ 1\ ]^T$?

According to Theorem 7.3, a Krylov solution only exists if and only if the right-hand side lies in the range of $\mathbf{W}^2$, since $\lambda = 0$ has index 2. This means that the right-hand side *must* be of the form $[c_1\ c_2\ 0\ 0]^T$ for some constants $c_1, c_2 \in \mathbb{R}$.

The minimal polynomial is seen to be $q(\lambda) = \lambda^2(\lambda - 2)$, so it has degree 3. According to Theorem 7.4, when a solution exists, it will lie in $\mathcal{K}_{3-2}(\mathbf{W}, \mathbf{b}) = \text{span}\{\mathbf{b}\}$.

So there is no Krylov solution to the system shown in Equation (D.14) and GMRES also answers with the message[4]

```
Matrix singular - no Krylov solution exists
ans =
    1.0000
    1.0000
         0
    0.5000
```

[4]The algorithm used is the one tailored by the author, see Section E.1 (the next section).

The solution returned was the best the algorithm could come up with, given the Krylov subspace available.

Note that if MATLAB 6.0's own `gmres` algorithm is used on this exact problem, we get the following output:

```
Warning: Divide by zero.
> In /appl/matlabr12/toolbox/matlab/sparfun/gmres.m at line 270
Warning: Divide by zero.
> In /appl/matlabr12/toolbox/matlab/sparfun/gmres.m at line 302
Warning: Matrix is singular to working precision.
> In /appl/matlabr12/toolbox/matlab/sparfun/gmres.m at line 317
gmres(4) stopped at iteration 1(4) without converging to the
desired tolerance 1e-06 because the maximum number of iterations
was reached.
The iterate returned (number 1(1)) has relative residual 0.33
ans =
    1.0000
    1.0000
         0
    0.5000
```

The approximate solution is the same as before, but the algorithm clearly divides with a number that is zero without hesitation. It finds out, though, after running the maximal number of iterations, that the solution found is not very good.

APPENDIX E

Source Code

This chapter will provide source code for a MATLAB implementation of GMRES. It should be seen in connection with Section 8.1 that deals with general implementation details of GMRES.

E.1 GMRES in MATLAB

```
function [x,res,X] = gmr(A,b,maxit,x0,tol)

%GMRES  Generalized Minimum Residual
%
% [x,res,X] = gmr(A,b,maxit,x0,tol)
%
% Input:  A     Square matrix
%         b     Right-hand side
%         maxit Maximum number of iterations (default: matrix
%               order)
%         x0    Starting guess
%         tol   Tolerance limit for the relative residual
% Output: x     (Approximate) solution
%         res   Vector of residuals at each iteration (found
%               implicitly, can be subject to rounding errors)
%         X     The approximate solution found after each iteration
%               is stored in each column

% Jan Marthedal Rasmussen, October 2000
%   Revised January 2001
```

```
% Check input arguments
[m,n] = size(A);
if m ~= n, error('A must be square'); end;
if size(b) ~= [n,1], error('b has wrong dimensions'); end;
if nargin < 2, error('Too few input arguments'); end;
if nargin < 3 | isempty(maxit), maxit = n; end;
if nargin < 4 | isempty(x0), x0 = zeros(size(b)); end;
if nargin < 5 | isempty(tol), tol = 2*eps; end;
maxit = min([n maxit]);

% Allocate space
res = zeros(1+maxit,1); % Residual vector
X   = zeros(n,1+maxit); % Matrix of solutions
V   = zeros(n,maxit);   % Orthon. vectors spanning the Krylov space
h   = zeros(maxit,1);   % New column of the upper Hessenberg
Q   = zeros(maxit+1);   % H = Q*T, Q orthogonal
T   = zeros(maxit);     %          T upper triangular
W   = zeros(n,maxit);   % W = V*inv(T)

% Initialize variables
r      = b - A*x0;
eta    = norm(r);
Q(1,1) = 1;
res(1) = eta;
x      = x0;
X(:,1) = x;

% Begin iterations
for k=1:maxit

  if eta <= tol*res(1), k=k-1; break; end;

  V(:,k) = r/eta;
  r = A*V(:,k);

  % Mod. Gram-Schmidt on the new vector
  for i=1:k
    h(i) = V(:,i)'*r;
    r = r - V(:,i)*h(i);
  end
  eta = norm(r);

  % Apply previous rotations to h
  T(1:k,k) = Q(1:k,1:k)'*h(1:k);

  % Compute Givens rotation parameters
  rc = T(k,k);

  if eta == 0
    c=1; s=0;
  elseif abs(eta) > abs(rc)
```



```
    tau = -rc/eta;
    s = 1 / sqrt(1 + abs(tau)^2);
    c = s*tau;
  else
    tau = -eta/rc;
    c = 1 / sqrt(1 + abs(tau)^2);
    s = c*tau;
  end

  % Apply Givens rotations
  T(k,k) = c'*rc - s'*eta;
  Q(1:k,[k k+1]) = Q(1:k,k)*[c s];
  Q(k+1,[k k+1]) = [-s c];

  if abs(T(k,k)) <= eps
    disp('Matrix (numerically) singular - no Krylov solution exists');
    k=k-1; break;
  end

  % Update W = V*inv(T)
  W(:,k) = (V(:,k) - W(:,1:k-1)*T(1:k-1,k))/T(k,k);

  % Update solution
  x = x + res(1)*Q(1,k)*W(:,k);

  % Update output variables
  res(k+1) = res(1)*abs(Q(1,k+1));
  X(:,k+1) = x;

end

% Cut of output variables if stopped prematurely
res = res(1:k+1);
X   = X(:,1:k+1);
```

Appendix F

Thesis Notes

For all experiments with GMRES, MATLAB version 6.0 was used. For all numerical experiments in general, the computers used double precision (64 bit floating point numbers) with unit roundoff $u \simeq 1.1 \cdot 10^{-16}$.

Numerical integration, as required by some of the examples in the appendix, was done by a C++ package called Cubpack++, see [CLP97]. To obtain the manual or to download the source code, see the URL http://www.cs.kuleuven.ac.be/~ronald/.

The kernels visualized on pages 36 and 72 were rendered using POV-Ray 3.1g for Windows, see http://www.povray.org. Setting up the scenes visually was done using Moray V3.2 For Windows, see http://www.stmuc.com/moray.

All the plots shown were made by producing data files using MATLAB and then using the LaTeX package PSTricks for the visual presentation.

To contact the author, e-mail at jmr@imm.dtu.dk. See also the homepage http://www.imm.dtu.dk/~jmr for other information that may be of relevance.

Bibliography

[Ans71] Philip M. Anselone. *Collectively Compact Operator Approximation Theory*. Prentice-Hall, 1971.

[Bak77] Christopher T. H. Baker. *The Numerical Treatment of Integral Equations*. Oxford University Press, 1977.

[CLP97] Ronald Cools, Dirk Laurie, and Luc Pluym. *A User Manual for Cubpack++, version 1.1*. Dept. of Computer Science, K. U. Leuven, February 1997.

[CLR00] D. Calvetti, B. Lewis, and L. Reichel. GMRES, L-curves, and discrete ill-posed problems. *BIT*, 48, 2000.

[Coc72] James Alan Cochran. *The Analysis of Linear Integral Equations*. McGraw–Hill Book Company, 1972.

[DS63] Nelson Dunford and Jacob T. Schwartz. *Linear Operators, Part II: Spectral Theory, Self Adjoint Opertors in Hilbert Space*. Interscience Publishers, 1963.

[DS64] Nelson Dunford and Jacob T. Schwartz. *Linear Operators, Part I: General Theory*. Interscience Publishers, 1964.

[Gre97] Anne Greenbaum. *Iterative Methods for Solving Linear Systems*. Frontiers in Applied Mathematics. SIAM, 1997.

[Gro93] Charles W. Groetsch. *Inverse Problems in the Mathematical Sciences*. Vieweg, 1993.

[GvL96] Gene H. Golub and Charles F. van Loan. *Matrix Computations*. The John Hopkins University Press, third edition, 1996.

[Hac95] Wolfgang Hackbusch. *Integral Equations, Theory and Numerical Treatment*. Birkhäuser, 1995.

[Han76] Erik Hansen. *Sædvanlige differentialligninger fra fysikken*. Polyteknisk forlag, Lyngby, second edition, 1976.

[Han88] P. C. Hansen. Computation of the Singular Value Expansion. *Computing*, 40:185–199, 1988.

[Han98a] Per Christian Hansen. *Rank-Deficient and Discrete Ill-Posed Problems*. SIAM, 1998.

[Han98b] Per Christian Hansen. Regularization tools. Technical Report IMM–REP–1998–6, IMM, DTU, 1998.

[Hoc73] Harry Hochstadt. *Integral Equations*. John Wiley and Sons, Inc., 1973.

[HS78] P. R. Halmos and V. S. Sunder. *Bounded Integral Operators on L^2 Spaces*. Springer-Verlag Berlin Heidelberg, 1978.

[HT31] E. Hille and J. D. Tamarkin. On the characteristic values of linear integral equations. *Acta Mathematica*, 57:1–76, 1931.

[IM98] Ilse C. F. Ipsen and Carl D. Meyer. The idea behind Krylov methods. *American Mathematical Monthly*, 105(10):889–899, December 1998. Available at `http://meyer.math.ncsu.edu`.

[Kar64] Samuel Karlin. The existence of eigenvalues for integral operators. *Transactions of the American Mathematical Society*, 113:1–17, 1964.

[Kar91] Rune Karlson. A study of some roundoff effects of the GMRES-method. Technical report, Department of Mathematics, Linköping University, Sweden, January 1991.

[Knu97] Donald E. Knuth. *The Art of Computer Programming*, volume 1 / Fundamental Algorithms. Addison–Wesley, third edition, 1997.

[Kre99] Rainer Kress. *Linear Integral Equations*, volume 82 of *Applied Mathematical Sciences*. Springer-Verlag, New York, second edition, 1999.

[Loa92] Charles Van Loan. *Computational Frameworks for the Fast Fourier Transform*. SIAM, 1992.

[LR84] G. Little and J. B. Reade. Eigenvalues of analytic kernels. *SIAM Journal of Mathematical Analysis*, 15(1):133–136, January 1984.

[MH87] Jerrold E. Marsden and Michael J. Hoffman. *Basic Compelex Analysis*. W. H. Freeman and Company, second edition, 1987.

[Nev93] Olavi Nevanlinna. *Convergence of Iterations for Linear Equations*. Birkhäuser Verlag, 1993.

[Ped00] Michael Pedersen. *Functional Analysis in Applied Mathematics and Engineering*. CRC Press LLC, 2000.

[Rud66] W. Rudin. *Real and Complex Analysis*. McGraw-Hill, 1966.

[Smi37] F. Smithies. The eigen-values and singular values of integral equations. *Proc. London Math. Soc.*, 43:255–279, 1937.

[SS86] Youcef Saad and Martin H. Schultz. GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems. *SIAM Sci. Stat. Comput.*, 7:856–869, 1986.

[SvdV93] G. L. G. Sleijpen and H. A. van der Vorst. Krylov subspace methods for large linear systems of equations. Technical Report Preprint 803, Department of Mathematics, University Utrecht, 1993.

[Swa71] Dale W. Swann. Kernels with only a finite number of characteristic values. *Proc. Camb. Phil. Soc.*, 70:257–262, 1971.

[Sze67] Gabor Szegö. *Orthogonal Polynomials*, volume 23 of *AMS Collequium Publications*. American Mathematical Society, Providence, Rhode Islands, third edition, 1967.

[Wey49] Hermann Weyl. Inequalities between the two kinds of eigenvalues of a linear transformation. *Proc. Nat. Acad. Sci.*, 35:408–411, 1949.

[You71] David M. Young. *Iterative Solution of Large Linear Systems*. Academic Press, 1971.