Student's Book
Numerical Functional Analysis

Editor: Stefan Engblom

2019

Preface

This portfolio collects the student's output during the course *Numerical Functional Analysis*, which was given for the second time in the spring 2019 at the Department of Information technology, Uppsala university.

The first part of the course went over the basics of Metric spaces, Normed spaces, and Inner product spaces, following closely the first three chapters of Kreyszig's book *Introductory functional analysis with applications*. The five 'Big' theorems of functional analysis were next presented by the students themselves: the Hahn-Banach theorem, the Uniform boundedness theorem, the Open mapping theorem, the Closed graph theorem, and the Banach fixed point theorem.

The final part of the course consisted of short essays on various topics, typically connecting Numerical analysis and Functional analysis in one way or the other. The essays were improved by double open review among the students themselves after which the final version entered this portfolio. A selection of student's solution to book exercises has also been included: the selection was very loosely based on the difficulty of the exercise and on the solution quality.

Stefan Engblom
Uppsala, September 2019

Contents

Part I

Student's essays

The Schauder-Tikhonov theorem and its applications

Anton Artemov*

August 28, 2019

Abstract

We consider a variant of a fixed-point theorem known as the Schauder-Tikhonov theorem. The proof for the Banach space case is given. The Tikhonov generalization to locally-convex sets is presented. Some applications in ODE theory and operator theory are given.

1 Auxiliary theory

In this section we state the auxiliary theory necessary for the proof of the Schauder-Tikhonov theorem. The theory is given without proofs.

1.1 Compactness

Theorem 1.1 (**Compactness criteria** [4]). *Let X be a Banach spance. Then the following three statements are equivalent:*

1. *$K \subset X$ is a compact set;*
2. *for any sequence $\{x_n\} \subset K, n \in \mathbb{N}$ there exists a subsequence $\{x_{n_k}\}, k \in \mathbb{N}$ such that the latter converges to $x \in K$;*
3. *K is closed and totally bounded, i.e. for any $\varepsilon > 0$ $\exists a_1, \ldots, a_n \in K : \ K \subset \bigcup_{i=1}^{N} B(a_i, \varepsilon)$, where $B(a, \varepsilon)$ is an open ball centered at a with radius ε.*

*Division of Scientific Computing, Department of Information Technology, Uppsala university, SE-751 05 Uppsala, Sweden. anton.artemov@it.uu.se

1.2 Nonlinear Schauder projector

Let X be a real Banach space, $K \subset X$ a compact subset, $f : K \longrightarrow K$ a continuous mapping.

According to Theorem 1.1 K is totally bounded, and thus contains a countable dense set $\{k_i, i \in \mathbb{N}\}$. For any $m \in \mathbb{N}$ there exists a finite subset $\{k_{i_l}, l = 1, \dots, M(m)\}$ such that

$$K \subset \bigcup_{i=1}^{M(m)} B(k_{i_l}, \frac{1}{m}). \tag{1.1}$$

For any $m \in \mathbb{N}$ one defines

$$\beta_l^m(x) = \max\left\{0, \frac{1}{m} - \|f(x) - k_{i_l}\|\right\},\ l = 1, \dots, M(m), x \in K. \tag{1.2}$$

One can show that that for any $x \in K$ there exists at least one l such that $\beta_l^m \neq 0$ [4]. This is important for the next object to be defined.

Definition 1.1. A mapping $f_m : K \longrightarrow \mathrm{span}\{k_{i_l}, l = 1, \dots, M(m)\}$ given by

$$f_m(x) = \frac{\sum\limits_{i=1}^{M(m)} \beta_l^m(x) k_{i_l}}{\sum\limits_{i=1}^{M(m)} \beta_l^m(x)} \tag{1.3}$$

is a **nonlinear Schauder projector**. It maps a compact set K to a convex hull generated by elements k_{i_l}.

The projector (1.3) has the following properties:

1. If f is continuous on K, then $\beta_l^m(x)$ is also continuous on K as max function is continuous and norm is continuous. Thus, the Schauder projector is continuous on K for any $m \in \mathbb{N}$.

2. If $\beta_l^m(x) \neq 0$, then $\|f(x) - k_{i_l}\| \leq \frac{1}{m}$. By the triangle inequality one obtains

$$\|f_m(x) - f(x)\| \leq \frac{\sum\limits_{i=1}^{M(m)} \beta_l^m(x) \|k_{i_l} - f(x)\|}{\sum\limits_{i=1}^{M(m)} \beta_l^m(x)} \leq \frac{1}{m}. \tag{1.4}$$

Thus $f_m \longrightarrow f$ uniformly on K as $m \longrightarrow +\infty$.

2 The Schauder-Tikhonov theorem

Theorem 2.1 (**Schauder**). *Let X be a real Banach space, $K \subset X$ a nonempty compact convex set, $f : K \longrightarrow K$ a continuous mapping. Then there exists a point $x_0 \in K$ such that $f(x_0) = x_0$.*

Proof. For any $m \in \mathbb{N}$ we construct the nonlinear Schauder projector f_m, which maps K to the closed convex hull K_m of $\{k_{i_l}, k = 1, \ldots, M(m)\}$.

Since K is convex, then $K_m \subset K$ and thus K_m is compact. Thus, if one constricts mapping f_m to $f_m : K_m \longrightarrow K_m$, it tuns out that it maps a compact convex subset of the finite-dimensional set $\mathrm{span}\{k_{i_l}, l = 1, \ldots, M(m)\}$ into itself. Since any normed finite-dimensional set is isomorphic to $\mathbb{R}^n$, then one can apply the Brouwer's fixed point theorem [3], and thus there exists a point $x_m \in K_m \subset K$ such that $f_m(x_m) = x_m$.

The sequence $\{x_m\}_{m\in\mathbb{N}}$ contains a subsequence $\{x_{m_j}\}_{j\in\mathbb{N}}$, which converges to some $x_0 \in K$ in a strong sense, because K is a compact set.

Then the following holds:

$$\begin{aligned} \|f(x_0) - x_0\| &= \|f(x_0) - f(x_{m_j}) + f(x_{m_j}) - x_{m_j} + x_{m_j} - x_0\| \\ &\leq \|f(x_0) - f(x_{m_j})\| + \|f(x_{m_j}) - x_{m_j}\| \\ &+ \|x_{m_j} - x_0\| \longrightarrow 0, \end{aligned} \tag{2.1}$$

since f is continuous, $f_{m_j}(x_{m_j}) = x_{m_j}$, $f_m \longrightarrow f$ uniformly on K and $x_{m_j} \longrightarrow x_0$ in the strong sense in K. This proves that $f(x_0) = x_0$.

□

The Brouwer's theorem handles the case of continuous function mapping a compact convex set to itself in $\mathbb{R}^n$. It serves as a basis for more general theorem like Schauder's or Tikhonov's versions.

2.1 History of the theorem

Juliusz Schauder proved the theorem for Banach spaces in 1930, and Theorem 2.1 is actually Schauder's version. Later, in 1935, Andrey Tikhonov made a non-trivial step and proved the result without completeness requirement [1]. His proof is more involved and is based on the notion of a locally-convex set. According to Tikhonov, a locally convex set is a topological vector space, such that there exists a basis of zero neighbourhoods (i.e. neighbourhoods of the zero element), which consist of convex sets. Any normed set is a locally convex set, the converse is not true. Tikhonov's generalization of the Schauder theorem is

Theorem 2.2 (**Tikhonov**)**.** *Any continuous mapping, which maps a compact convex subset of a locally-convex set to itself, has at least one fixed point.*

3 Applications

3.1 Differential equations

One of the earliest application concerns solvability of countable quasi-linear ODE systems. In [5] the following result is stated: the system

$$\begin{cases} \frac{dy_\alpha}{dx} &= f_\alpha(x, \ldots, y_\beta, \ldots) \\ y_\alpha(x_0) &= y_\alpha^0 \end{cases}, \alpha, \beta \in \mathfrak{M}, \tag{3.1}$$

where f is a bounded continuous function of finitely (or countably many) variables, where the continuity for countably many arguments means continuity for any finite number of arguments, α and β characterize cardinality of those sets, has a solution. The proof of this result relies on the Schauder-Tikhonov theorem and constructs the integral representation of the system (3.1), which turns out to be a mapping that maps a convex compact set of functions to itself, thus has a fixed point.

3.2 Invariant subspaces

The problem of finding invariant subspaces is often referred to as *the* problem of operator theory [3, p. 46]. Given a Banach space X and a bounded linear operator T on X, one needs to find a closed subspace $M \subset X$ such that $M \neq X, M \neq \emptyset$ and $TM \subset M$. If such M exists, then it is called an invariant subspace for T. Not all bounded linear operators on Banach spaces have invariant subspaces. One can also define a *hyperinvariant* subspace for T : it is a subspace invariant for all operators commuting with T. Lomonosov [2] uses the Schauder-Tikhonov theorem to prove the following most general result:

Theorem 3.1 (**Lomonosov**)**.** *Let X be a Banach space, T be a non-scalar (not a multiple of identity) bounded linear operator commuting with a nonzero compact operator S, then T has a hyperinvariant subspace.*

The proof, which is based on a contradiction, assumes that T lacks eigenvalues and thus has no hyperinvariant spaces. Then, a special mapping is constructed, which is continuous and maps a compact convex set (a closed ball) to itself, thus has a fixed point. The existence of a fixed point is used to show that T does have an eigenvector.

References

[1] E. M. Bogatov. On the history of the fixed point method and the contribution of the Soviet mathematicians, (1920s-1950s.). *Chebyshevskii sbornik*, 19(2):30–55, 2018.

[2] V. I. Lomonosov. Invariant subspaces for the family of operators which commute with a completely continuous operator. *Functional Analysis and Its Applications*, 7(3):213–214, Jul 1973.

[3] Vittorino Pata. Fixed point theorems and applications. *Politecnico di Milano*, 2014.

[4] S. A. Sazhenkov. *Lecture notes on applied functional analysis*. Novosibirsk State University, 2013.

[5] Andrei Tychonoff. Ein Fixpunktsatz. *Mathematische Annalen*, 111(1):767–776, 1935.

Integral equations for Stokes flow

Joar Bagge*

September 2, 2019

Abstract

We consider the equation $(A - \lambda I)\varphi = f$ in a normed space and present the *Fredholm alternative*, which links the existence and uniqueness of the solution φ to compactness of the operator A. The case where A is an integral operator is of special interest, and we give criteria for when such an integral operator is compact.

We apply the theory to Stokes flow. We show how the equations of Stokes flow can be reformulated as an integral equation, and that the corresponding integral operator satisfies the criteria for compactness. Thus the Fredholm alternative can be applied to Stokes flow.

1 Introduction

Stokes flow is a linearized model of an incompressible fluid. It is valid when the characteristic speed or length scale is very small, for example in microscopic flows (see for example Pozrikidis [1992] or Kim and Karrila [1991]). In the absence of external forces, the fluid flow is governed by the *Stokes equations*

$$\begin{aligned} \nabla p - \nabla^2 u &= 0, \\ \nabla \cdot u &= 0. \end{aligned} \tag{1.1}$$

Here, u is the velocity of the fluid (a vector field) and p is the pressure (a scalar field).[1]

*Division of Numerical Analysis, Department of Mathematics, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. joarb@kth.se

[1]The first equation in (1.1) is actually Newton's second law, expressing that the momentum of the fluid is conserved. The second equation in (1.1) expresses that the fluid is incompressible, and that its mass is conserved. The equations have been *nondimensionalized* so that irrelevant constants disappear.

The linearity of the Stokes equations (1.1) makes them possible to reformulate as an integral equation of the form

$$(A - \lambda I)\varphi = f, \tag{1.2}$$

where the unknown φ and the data f are elements in some normed space X, $A : X \to X$ is an operator, $I : X \to X$ is the identity operator (i.e. $I\varphi = \varphi$ for all $\varphi \in X$) and $\lambda \neq 0$ is a real (or complex) number.

In section 2, we study the existence and uniqueness of solutions to the general equation (1.2) in terms of compactness of A. In section 3, we describe how (1.1) can be reformulated as an integral equation of the form (1.2). In section 4, we show that the integral operator associated with Stokes flow is compact if Γ is a Lyapunov surface.

2 Compact integral operators

A crucial property of the operator A in (1.2) is *compactness*; if A is a *compact operator* then the existence and uniqueness of a solution φ to (1.2) can be characterized using the so-called *Fredholm alternative.* Before we define compactness of an operator, recall that a *subset* V of a metric space is called compact if every sequence in V has a convergent subsequence whose limit is also an element of V.

Definition 2.1. A linear operator $A : X \to Y$, where X and Y are normed spaces, is called *compact* if for every bounded subset U of X, the image $A(U)$ is *relatively compact*, i.e. its closure $\overline{A(U)}$ is a compact subset of Y.

Theorem 2.1 (Fredholm alternative)**.** *Let $A : X \to X$ be a compact linear operator on a normed space X, and let $\lambda \neq 0$ be a real (or complex) number. Then exactly one of the following holds.*

(I) The equation $(A - \lambda I)\varphi = f$ has a unique solution $\varphi \in X$ for every given $f \in X$.

(II) The homogeneous equation $(A - \lambda I)\varphi = 0$ has a nontrivial solution $\varphi \neq 0$.

The proof of Theorem 2.1 is one of the main highlights of chapter 8 in Kreyszig [1978], to which we refer the interested reader. The significance of Theorem 2.1 is that in order to prove that the nonhomogeneous equation (1.2) has a unique solution for every $f \in X$, it suffices to show that the homogeneous equation has only the trivial solution.

Theorem 2.1 is general and holds for any compact linear operator A on a normed space. However, of special interest are *integral operators*, which take a function and convolve it with a kernel. More specifically, for an integer $m \geq 2$, let $\Omega \subseteq \mathbb{R}^m$ be a bounded open domain with boundary Γ. We consider an integral operator $A : C(\Gamma) \to C(\Gamma)$ on the space of continuous functions on Γ, given by

$$(A\varphi)(x) = \int_\Gamma K(x, y)\varphi(y)\, dS_y, \qquad x \in \Gamma, \quad \varphi \in C(\Gamma), \tag{2.1}$$

where K is the kernel, defined everywhere on $\Gamma \times \Gamma$ except possibly on the set $\{(x, y) \in \Gamma \times \Gamma : x = y\}$ where it may be singular.

Definition 2.2. The kernel K in (2.1) is called *weakly singular* if it is defined and continuous for all $x, y \in \Gamma$ such that $x \neq y$, and there exist constants $M > 0$ and $\alpha \in (0, m - 1]$ such that

$$|K(x, y)| \leq M|x - y|^{\alpha-m+1}, \qquad x, y \in \Gamma, \quad x \neq y.$$

Note that the term "weakly singular" depends on the dimension m of the space $\mathbb{R}^m$ in which Γ lies.

The following theorem from Kress [2014], p. 31, is the central theorem of this essay. It gives a sufficient condition for an integral operator to be a compact operator. We give a sketch of the proof; for the full proof, see Kress [2014], pp. 28–32.

Theorem 2.2. *Let the integral operator A be defined by (2.1) with weakly singular kernel. Then A is a compact operator if Γ is of class C^1.*

Proof sketch. The proof consists of three steps:

1. Verify that the integral defined by (2.1) with weakly singular kernel exists as an improper integral for all $x \in \Gamma$. To do this, note that since Γ is C^1, the normal vector is continuous on Γ. Consider a small neighbourhood $S(x; R) := \{y \in \Gamma : |y - x| \leq R\}$ and show that the integral
$$\int_{S(x;R)} K(x, y)\varphi(y)\, dS_y$$
can be bounded by using polar coordinates in the tangent plane at x. Thus (2.1) exists.

2. Prove that (2.1) defines a compact operator in the special case when the kernel is continuous everywhere on $\Gamma \times \Gamma$. To do this, take any bounded

subset U of $C(\Gamma)$. Show that $A(U)$ is bounded. Since K is continuous on the compact set $\Gamma \times \Gamma$, it is uniformly continuous. From this it follows that $A(U)$ is equicontinuous[2]. By the Arzelà–Ascoli theorem (Theorem 1.18 in Kress [2014]), $A(U)$ is relatively compact. Thus A is a compact operator in the special case when the kernel is continuous.

3. Finish the proof in the general case by defining a sequence of kernels K_n which are continuous everywhere on $\Gamma \times \Gamma$ and converge to the weakly singular K as $n \to \infty$. By step 2, the corresponding integral operators A_n are compact. Show that $A_n\varphi \to A\varphi$ uniformly as $n \to \infty$, and that $||A_n - A||_\infty \to 0$ as $n \to \infty$. By Theorem 2.22 in Kress [2014], A is then a compact operator. $\square$

3 The Stokes integral formulation

We now go on to present how the Stokes equations (1.1) can be reformulated as an integral equation of the form (1.2), based on Pozrikidis [1992] and Kim and Karrila [1991].

Let $\Omega \subseteq \mathbb{R}^3$ ($m = 3$) be a bounded open domain with boundary Γ. We consider the *interior Dirichlet problem*, in which the velocity u is known on Γ, say

$$u(x) = g(x), \qquad x \in \Gamma, \tag{3.1}$$

where $g : \Gamma \to \mathbb{R}^3$ is known, and we seek the solution u to (1.1) in the domain Ω. (The pressure p is also unknown, but it is not of much interest to us here, and mainly serves to ensure that the incompressibility condition $\nabla \cdot u = 0$ can be fulfilled.)

In a *boundary integral formulation*, the velocity u is expressed as an integral over the boundary Γ, for example as[3]

$$u(x) = (Dq)(x) := \int_\Gamma K(x,y)q(y)\,dS_y, \qquad x \in \Omega, \tag{3.2}$$

where $D : C(\Gamma)^3 \to C(\Omega)^3$ is an integral operator known as the *double layer operator*, which takes a continuous vector field q defined on Γ and returns a continuous vector field Dq defined on Ω. Note that D is not of the form of

[2] A subset $F \subseteq C(\Gamma)$ is called (uniformly) *equicontinuous* if for every $\varepsilon > 0$ there exists $\delta > 0$ such that for all $\varphi \in F$ it holds that $|\varphi(x) - \varphi(y)| < \varepsilon$ whenever $|x - y| < \delta$, where $x, y \in \Gamma$. Note that δ is independent of φ, x and y.

[3] It can be shown that for any $q \in C(\Gamma)^3$, the function Dq defined by (3.2) is actually a solution to (1.1) in Ω for some pressure field p. (Pozrikidis [1992], Kim and Karrila [1991])

the operator A in (2.1). The *kernel* $K : \Omega \times \Gamma \to \mathbb{R}^{3\times 3}$ is here a matrix-valued function given by

$$K_{ij}(x, y) = \sum_{k=1}^{3} T_{ijk}(y - x) n_k(y), \tag{3.3}$$

where the *stresslet* T is given by

$$T_{ijk}(r) = -6\frac{r_i r_j r_k}{|r|^5}, \qquad r \in \mathbb{R}^3$$

and n is the outward-pointing unit normal of Γ (which we assume to be defined at every point of Γ). We are interested in the operator D_Γ which corresponds to D in (3.2) when x is restricted to Γ. The operator D_Γ is defined as follows.

Definition 3.1. Let the integral operator $D_\Gamma : C(\Gamma)^3 \to C(\Gamma)^3$ be given by

$$(D_\Gamma q)(x) := \int_\Gamma K(x, y) q(y)\, dS_y, \qquad x \in \Gamma, \tag{3.4}$$

where the kernel K is given by (3.3) and defined for all $(x, y) \in \Gamma \times \Gamma$ except on the line $x = y$, where it is singular.

The two operators D and D_Γ are related by a jump discontinuity: if $x \in \Gamma$ and $z \in \Omega$ then

$$\lim_{z\to x} (Dq)(z) = (D_\Gamma q)(x) - 4\pi q(x). \tag{3.5}$$

The goal is to find the function q such that the velocity u resulting from (3.2) solves (1.1) and satisfies the boundary condition $u(x) = g(x)$ for $x \in \Gamma$. The velocity should be continuous, so $u(x) = \lim_{z\to x} u(z) = \lim_{z\to x} (Dq)(z)$, and using (3.5) we arrive at the equation $(D_\Gamma q)(x) - 4\pi q(x) = g(x)$, $x \in \Gamma$, which can be written as

$$(D_\Gamma - 4\pi I)q = g, \tag{3.6}$$

where I is the identity operator. This equation is of the form (1.2), and D_Γ given by (3.4) is of the form (2.1) with the exception that it acts on vector-valued functions and has a matrix-valued kernel. However, these differences can be handled by considering each component separately.

4 Compactness of the integral operator D_Γ

We now show that D_Γ is a compact integral operator if Γ is a so-called Lyapunov surface. This requirement on Γ is needed to ensure that the kernel of D_Γ is weakly singular, so that Theorem 2.2 is applicable. The following definition is due to Kim and Karrila [1991], p. 24.

Definition 4.1. A surface Γ in $\mathbb{R}^3$ is called *Lyapunov* if it is of class C^1 and there are constants $M' > 0$ and $\beta \in (0, 1]$ such that

$$|n(y) \cdot (y - x)| \leq M'|y - x|^{1+\beta}$$

for all $x, y \in \Gamma$ such that $x \neq y$, where n is a unit normal of Γ.

Theorem 4.1. *Let Γ be the boundary of a bounded open domain $\Omega \subseteq \mathbb{R}^3$. The integral operator D_Γ given by (3.4) is a compact operator from $C(\Gamma)^3$ to $C(\Gamma)^3$ if Γ is a Lyapunov surface.*

Proof. We can rewrite (3.4) more explicitly as

$$(D_\Gamma q)_i(x) = \sum_{j=1}^{3} \int_\Gamma K_{ij}(x, y) q_j(y)\, dS_y, \qquad x \in \Gamma. \tag{4.1}$$

We claim that D_Γ is compact if each of its components is compact, and that a sum of compact operators is itself compact. Thus, it is enough to show that $\int_\Gamma K_{ij}(x, y)q_j(y)\, dS_y$ is a compact integral operator, which by Theorem 2.2 holds if the kernel K_{ij} is weakly singular. Let us verify this. Clearly, K_{ij} as given by (3.3) is continuous everywhere except on $x = y$. This is because the normal n is continuous on Γ since the latter is Lyapunov and therefore C^1. Furthermore, writing $r = y - x$, we have by definition of the kernel

$$|K_{ij}(x, y)| = \left| \sum_{k=1}^{3} \left(-6 \frac{r_i r_j r_k}{|r|^5} \right) n_k(y) \right| = \left| 6 \frac{r_i r_j}{|r|^5} \right| \cdot \underbrace{\left| \sum_{k=1}^{3} r_k n_k(y) \right|}_{\text{(A)}}.$$

The part labelled (A) is $|r \cdot n(y)|$, which is bounded by $M'|r|^{1+\beta}$ since Γ is Lyapunov (Definition 4.1). Together with the inequality $|r_i| \leq |r|$, $i = 1, 2, 3$, this implies that there are constants $M' > 0$ and $\beta \in (0, 1]$ such that

$$|K_{ij}(x, y)| \leq 6 \frac{|r|^2}{|r|^5} \cdot M'|r|^{1+\beta} = M|r|^{\beta - 3 + 1}$$

for all $x, y \in \Gamma$ such that $x \neq y$. Since $m = 3$, this shows that K_{ij} is weakly singular, with $M = 6M'$. By Theorem 2.2, the operator D_Γ is compact. $\square$

To summarize, we have shown that the Stokes integral operator D_Γ is compact, which means that the Fredholm alternative (Theorem 2.1) can be applied to the integral equation (3.6). Thus, existence and uniqueness of a solution can be guaranteed for any continuous Dirichlet boundary condition, by showing that the homogeneous case has only the trivial solution.

References

S. Kim and S. J. Karrila. *Microhydrodynamics: principles and selected applications*. Butterworth-Heinemann, Boston, 1991.

R. Kress. *Linear Integral Equations*. Springer, New York, 3rd edition, 2014.

E. Kreyszig. *Introductory functional analysis with applications*. Wiley, New York, 1978.

C. Pozrikidis. *Boundary integral and singularity methods for linearized viscous flow*. Cambridge University Press, Cambridge, 1992.

Fredholm Alternative: a short introduction

Samuel Bronstein*

September 6, 2019

Abstract

In this report we present a proof of the Fredholm Alternative. It is a very general theorem in Functional Analysis, with applications in multiple domains, and especially in integral equations. Fredholm Operators are strongly associated with Integral Operators as well.

1 Introduction

The Fredholm Alternative is a theorem by Ivar Fredholm (1866-1927), a mathematician who has had a major impact on Spectral Theory. For a more detailed description see Steen [1973].

This result is especially useful when looking at the solutions of an equation of the type $x - Tx = y$ for $y \in E$ and $T : E \to E$ an operator. This application can be found in Fredholm [1903], or more clearly in [Kress et al., 1989, corollary 4.18]

At first, we state the theorem when T is a compact operator, as in that case the proof is quite simple. From this theorem one extracts the notion of Fredholm operators, and there is a lot to say about those operators, as we will give some examples in the last section.

1.1 Compact Operators

Let's introduce first the notion of compact operator.

Definition 1.1. Let E and F be two Banach spaces. We denote $B(E)$ the unit ball of E. An operator $T \in \mathcal{L}(E, F)$ is said to be **a compact operator** if the image of $B(E)$ under T has compact closure in F.

*Department of Mathematics, Ecole Normale Superieure, 75005 Paris, France. samuel.bronstein@ens.fr

The notion of compact operators comes quite naturally as a generalization of an operator of finite rank. Instead of asking the image of the unit ball to be a bounded subset of a finite dimentional space, we ask it to be a subset of a compact set.

Remark. As we deal with bounded operators, if E or F is finite dimensional, compact operators are of finite rank. If F is a Hilbert space, then compact operators are the closure of bounded operators of finite rank. For more details, see the work of Brézis [1983].

2 The main theorem

The Fredholm Alternative concerns the solutions of an equation of the type $x - Tx = y$ where $y \in E$ and T is a compact endomorphism. We will denote $T^\star$ the adjoint operator of T. This alternative is actually very useful when trying to invert the operator $I - T$, as it can be formulated in this way, close to Fredholm's original work Fredholm [1903].

- If $I - T$ is bijective, for any $y \in E$ there is a unique solution x to the equation $x - Tx = y$.
- If $I - T$ is not bijective, there is an integer n and n conditions of orthogonality such that:
  - if y satisfies the conditions, the set of solutions to $x - Tx = y$ is a n dimensional affine subset
  - if y doesn't satisfy the conditions, there is no solution to the equation.

We will prove a more modern formulation, which is the following.

Theorem 2.1. *Let E be a Banach space, I be the identity map of E and T a compact endomorphism of E. Then we have the following properties.*

1. *The kernel* $\ker(I - T)$ *is finite dimensional.*
2. *The image* $\mathrm{im}\,(I - T)$ *is closed and is equal to the orthogonal of* $\ker(I - T^\star)$. [1]
3. *$I - T$ is injective if and only if it is surjective.*

[1] If A is a subset of an inner product space E, the orthogonal of A is the set of elements y that are orthogonal to all elements of A. It is denoted $A^\perp$.

4. $\dim\ker(I - T) = \dim\ker(I - T^\star)$*.*

Remark. A nice property of compact operators is that for λ a scalar and T a compact operator, λT is a compact operator.
This means that the discussion given for the equation $x - Tx = y$ can also be given for $\lambda x - Tx = y$, where λ is a nonzero scalar.

3 Proof

In order to make the proof easier, we will use the following lemmas.

Lemma 3.1. *Let M be a subspace of E a Banach space. We have $(M^\perp)^\perp = \overline{M}$.*

Proof. It is obvious that $\overline{M} \subset (M^\perp)^\perp$. For the other inclusion, we use the Hahn-Banach theorem. If x doesn't belong to $\overline{M}$, we know that there is a linear form f and $\varepsilon > 0$ such that $\overline{M} \subset \ker(f)$ and $f(x) > \varepsilon$. □

Lemma 3.2. *Let E be a normed vector space. If M is a strict closed subspace of E then for all $\varepsilon > 0$ there is $x \in E$ such that:*

$$\begin{cases} ||x|| = 1 \\ d(x, M) \geq 1 - \varepsilon \end{cases}$$

Proof. Let y be in E but not in M. As M is closed, $d(y, M) > 0$.
So there is $y' \in M$ such that

$$d(y, M) \leq ||y - y'|| \leq \frac{d(y, M)}{1 - \varepsilon}$$

Let $z \in M$ and

$$x = \frac{y - y'}{||y - y'||}$$

Let's show that $||x - z|| \geq 1 - \varepsilon$. We have

$$\begin{aligned} x - z &= \frac{y - y'}{||y - y'||} - z \\ &= \frac{1}{||y - y'||}(y - y' - z||y - y'||) \\ &= \frac{1}{||y - y'||}(y - (y' + z||y - y'||)) \end{aligned}$$

As $y' + z||y - y'||$ belongs to M we deduce that $||x - z|| \geq 1 - \varepsilon$. □

- For the first statement, let K_0 be $\ker(I-T)$ and $B(K_0)$ its unit ball. As $B(K_0) = T(B(K_0))$ is a subset of $T(B(E))$, it is compact. So, by the Riesz characterization of finite dimensional spaces, one has that K_0 is finite dimensional.

- For the second one, the first inclusion $\mathrm{im}\,(I-T) \subset \ker(I-T^\star)^\perp$ is clear, as for any $y = x - Tx$ and $z \in \ker(I-T^\star)$. We have:

$$\begin{aligned}(y,z) &= (x-Tx,z)\\ &= (x,z)-(x,T^\star z)\\ &= (x,z)-(x,z)\\ &= 0\end{aligned} \tag{3.1}$$

From this one deduces that $\overline{\mathrm{im}\,(I-T)} \subset \ker(I-T^\star)^\perp$. We know that, by definition, $\ker(I-T^\star) = \mathrm{im}\,(I-T)^\perp$. Therefore, thanks to lemma 3.1, we just have to show the closedness of $\mathrm{im}\,(I-T)$.
Let (y_n) be a sequence belonging to $\mathrm{im}\,(I-T)$ such that $y_n \to y$. We have that $y_n = x_n - Tx_n$ for all n. Let $d_n = d(x_n, \ker(I-T))$. At first, we show that the sequence (d_n) is bounded. Suppose that there is a subsequence such that $d_n \to \infty$. As $\ker(I-T)$ is finite dimensional, for all n there is x'_n such that $d_n = ||x_n - x'_n||$. Then, noting $z_n = \frac{x_n - x'_n}{d_n}$ we have:

$$d(z_n, \ker(I-T)) = \frac{1}{d_n} d(x_n, \ker(I-T)) = 1 \tag{3.2}$$

and

$$z_n - Tz_n = \frac{y_n}{d_n} \to 0 \tag{3.3}$$

As T is a compact operator, we can extract again a subsequence and have a z such that $Tz_n \to z$. And as $z_n - Tz_n \to 0$, we deduce $z_n \to z$. This means that $z \in \ker(I-T)$ and $d(z, \ker(I-T) = 1$, which is absurd. So we know that (d_n) is bounded. The boundedness means that we can extract a subsequence and have x such that $T(x_n - x'_n) \to x$. As $y_n = x_n - x'_n - T(x_n - x'_n)$, we deduce that $x_n - x'_n \to y + x$.
Finally, this means that $y = (y+x) - T(y+x)$, finishing the proof of the second statement.

- For this item, we will only show that $I-T$ cannot be injective and not surjective. The other statement left to proove is gotten the same way via the adjoints.
So, suppose that $I-T$ is injective and not surjective.

- We denote (E_n) the sequence $E_n = (I-T)^n(E)$. As $I-T$ is not surjective, E_1 is distinct from E. And E_1 is stable under T. As T is injective and E_1 is distinct from E, we must have E_2 distinct from E_1. A recurrence shows that E_{n+1} is a propper subspace of E_n. The second statement then lets us state that the restriction of T on E_n is compact (as the images are closed). This means that (E_n) is a strictly decreasing sequence of closed subspaces.
- Using the lemmas, we build a sequence (x_n) such that $x_n \in E_n$, $||x_n|| = 1$ and $d(x_n, E_{n+1}) \geq \frac{1}{2}$. Then having $n > m$, we get that

  $$T(x_n - x_m) = \big((x_m - Tx_m) - (x_n - Tx_n) + x_n\big) - x_m \tag{3.4}$$

  and as the term in the parentheses belongs to E_{m+1}, we have

  $$||T(x_n - x_m)|| \geq \frac{1}{2} \tag{3.5}$$

- The compactness of T makes this impossible. So if $I-T$ is injective, $I-T$ is surjective.

We already said that the reciprocal way can be gotten with the same reasoning on $T^\star$.

- For the last statement, let's denote $d = \dim\ker(I-T)$ and $d^\star = \dim\ker(I-T^\star)$ and suppose that $d < d^\star$. As $\ker(I-T)$ is finite dimensional, we know that there is F such that $E = \ker(I-T)\bigoplus F$ and the projections are continuous.
Let p be one continuous projection on $\ker(I-T)$. As $\mathrm{im}\,(I-T) = \ker(I-T^\star)$ is closed, it has a topological supplementary $\tilde{E}$ of dimension $d^\star$. As $d < d^\star$, there exists a linear map $l : \ker(I-T) \mapsto \tilde{E}$ which is injective but not surjective.
Let's denote finally:

$$B = T + l \circ p \tag{3.6}$$

As $l \circ p$ is of finite rank, B is a compact operator also.
B is also built such that $I-B$ is injective, and applying the third statement we get that $I-B$ is surjective.
But this is impossible, as for a $y \in \tilde{E} \backslash \mathrm{im}\ l$ the equation $x - Bx = y$ does not have a solution. So $d \geq d^\star$.
Applying the same results to the operator $T^\star$, one gets that :

$$\dim\ker(I-T^{\star\star}) \leq \dim\ker(I-T^\star) \leq \dim\ker(I-T) \tag{3.7}$$

But it is easy to show that $\ker(I-T) \subset \ker(I-T^{\star\star})$, and so $d = d^\star$. This finishes the proof.

4 Miscellaneous

A natural way to go further is to define a Fredholm operator in the following way:

Definition 4.1. An operator A is called Fredholm if and only if: im A and im $A^\star$ are closed and.

$$\dim \ker A = \dim \ker A^\star < \infty \tag{4.1}$$

This way, the theorem proven above states that if T is a compact operator, then $I + T$ is a Fredholm operator.
Alexander Ramm gave in 2001 a characterization of Fredholm operators in Hilbert spaces. He showed the following theorem.Ramm [2001]

Theorem 4.1. *In a Hilbert space H, an operator A is a Fredholm operator if and only if $A = B + F$, where B is an isomorphism and F is a finite rank operator.*

References

H. Brézis. Analyse fonctionnelle, théorie et applications,(1983), 1983.

I. Fredholm. Sur une classe d'équations fonctionnelles. *Acta mathematica*, 27(1):365–390, 1903.

R. Kress, V. Maz'ya, and V. Kozlov. *Linear integral equations*, volume 82. Springer, 1989.

A. G. Ramm. A simple proof of the Fredholm alternative and a characterization of the Fredholm operators. *The American Mathematical Monthly*, 108(9):855–860, 2001.

L. A. Steen. Highlights in the history of spectral theory. *The American Mathematical Monthly*, 80(4):359–381, 1973.

Optimal control problems with state constraints for systems governed by differential inclusions

Ivo Dravins*

September 2, 2019

Abstract

This work aims to summarize some results from the article *On the generalized Bolza problem and its application to dynamic optimization* by A. D. Ioffe [2019]. We present two theorems from the article combined with the necessary notation and concepts. Finally, we speculate briefly on what the tools deployed by Ioffe may tell us in the context of PDE-constrained optimization.

1 Subdifferentials and preliminaries

Subdifferentials (or subderivatives) generalize the concept of the derivative to convex functions which are not necessarily differentiable (see Boyd and Vandenberghe [2008] for an introduction or Messaoud [2012] for a more comprehensive treatment). They are a tool often deployed in the study of convex functions, often within the context of convex optimization. In the paper by Ioffe several types of sub-differentials are used, they are:

- Proximal and limiting subdifferentials in in $\mathbb{R}^n$ and general Hilbert spaces. Denoted ∂_p and ∂ respectively.
- Dini-Hadamard subdifferentials in separable Banach spaces. Denoted ∂^-.
- G-subdifferentials in Banach spaces. Denoted ∂_G.

*Division of Scientific Computing, Department of Information Technology, Uppsala university, SE-751 05 Uppsala, Sweden. ivo.dravins@it.uu.se

- Clarke's generalized gradients in Banach spaces. Denoted ∂_C.

Throughout the work $B(x,r)$ denotes the closed ball of radius r around x and B denotes the unit ball, furthermore $\text{conv}(\cdot)$ denotes the convex hull.

2 Main Result

We consider the generalized Bolza problem:

$$\text{Minimize } J(x(\cdot)) = \varphi(x(\cdot)) + \int_0^T L(t, x(t), \dot{x}(t))dt$$

with φ a function on $C[0,T]$. We fix some $\bar{x}(\cdot) \in W^{1,1}$ which will be assumed to be a local minimizer of J. We list the basic assumptions made on the functions φ and L:

- (A_1) φ is Lipschitz in a neighborhood of $\bar{x}(\cdot)$ in $C[0,T]$, that is $\exists K > 0$ and $\bar{\varepsilon} > 0$ such that:

  $$|\varphi(x(\cdot)) - \varphi(x'(\cdot))| \leq K\|x(\cdot) - x'(\cdot)\|$$

  if $x(\cdot),\ x'(\cdot)$ are $\bar{\varepsilon}$-close to $\bar{x}(\cdot)$ in $C[0,T]$.

- (A_2) $L(t,x,y)$ is lower semicontinuos in (x,y) and measurable. Further there are $\bar{\varepsilon} > 0$, measurable $r(t) > 0$, summable $R(t) \geq 0$ and $k(t) \geq 0$ and a measurable set-valued mapping $D(\cdot) : [0,T] \rightrightarrows \mathbb{R}^n$ such that for almost every $t \in [0,T]$ the function $L(t,\cdot)$ is continuous on $B(\bar{x}(t), \bar{\varepsilon}) \times D(t)$ and

  $$|L(t,x,y) - L(t,x',y)| \leq R(t)\|x - x'\|,\ \forall x, x' \in B(\bar{x}(t), \bar{\varepsilon}), y \in B(\dot{\bar{x}}(t), r(t));$$

  $$|L(t,x,y)| \leq k(t),\ \forall x \in B(\bar{x}(t), \bar{\varepsilon}),\ y \in B(\dot{\bar{x}}(t), \bar{\varepsilon}).$$

  hold almost everywhere on $[0,T]$.

Theorem 1

Assume (A_1) and (A_2), If $\bar{x}(\cdot)$ is a local minimizer of J on $W^{1,1}$, then there is an $\mathbb{R}^n$-valued Radon measure $\nu \in \partial_G \varphi(\bar{x}(\cdot))$ (see Rad), an $\mathbb{R}^n$-valued function of bounded variations $p(t)$, continuous from the left and a summable $\mathbb{R}^n$-valued function $q(t)$ such that:

$$p(t) = -\int_t^T g(s)ds - \int_t^T d\nu(t), \quad p(0) = 0,$$

and the following conditions are satisfied almost everywhere on $[0,T]$:

$$q(t) \in \mathrm{conv}\{q : (q,p(t)) \in \partial L(t,\cdot,\cdot)(\bar{x}(t),\dot{\bar{x}}(t))\} \ ; \ \textit{Euler inclusion}$$

and

$$L(t,\bar{x}(t),u) - L(t,\bar{x}(t),\dot{\bar{x}}(t)) - \langle p(t), u-\dot{\bar{x}}(t)\rangle \geq 0 \text{ if } u \in D(t) \ ; \\ \textit{Weierstrass condition}$$

We can deduce from this theorem that the Euler inclusion and the Weierstrass condition together give a necessary condition for a strong minimum in the classical sense, that is when we have $\varepsilon > 0$ such that $J(x(\cdot)) \geq J(\bar{x}(\cdot)) \ ; \ \forall x(\cdot) \in W^{1,1}$ satisfying $\|x(t)-\bar{x}(t)\| < \epsilon \ \forall t$.

3 Optimal control application

We consider the optimal control problem:

$$\text{Minimize } l(x(0),x(T)) \tag{3.1}$$

$$\text{such that } \dot{x} \in F(t,x), \ (x(0),x(T)) \in S, \ g(t,x(t)) \leq 0, \forall t \tag{3.2}$$

where we assume that there are $\bar{x}(\cdot) \in W^{1,1}$ and $\bar{\varepsilon} > 0$ such that:

- (A_3) l is Lipschitz near $(\bar{x}(0),\bar{x}(T))$, $S \subset \mathbb{R}^n \times \mathbb{R}^n$ is closed;
- (A_4) g is upper semicontinuous on the set $\{(t,x) : t \in [0,T], \|x-\bar{x}(t)\| \leq \bar{\varepsilon}\}$ and there is a $K > 0$ such that for $x, x' \in B(\bar{x}(t),\bar{\varepsilon})$:

  $$|g(t,x) - g(t,x')| \leq K\|x-x'\| \text{ a.e. on } [0,T].$$

- (A_5) F is closed-valued and measurable in the standard sense and there is a measurable $\bar{r}(t) > 0$ bounded away from zero, a summable $\bar{R}(t) \leq 0$ and an $\eta \in (0,1)$ such that the relations

  $$F(t,x) \cap B(\dot{\bar{x}}(t),\bar{r}(t)) \subset F(t,x') + \bar{R}(t)\|x-x'\|B,$$
  $$F(t,x) \cap B(\dot{\bar{x}}(t),(1-\eta)\bar{r}(t)) \neq \emptyset$$

  hold $\forall x, x' \in B(\bar{x}(t),\bar{\varepsilon})$.

The first of these relations is a non-local extension of Aubin's pseudo-Lipschitz property (see Klatte and Kummer [2013]). We need some more notation before we are ready to state the theorem: Let $U(t)$ be the closure of the collection of all $u \in F(t, \bar{x}(t))$ such that:

$F(t,x) \cap B(u,r) \subset F(t,x') + R\|x - x'\|B$ and $F(t,x) \cap B(u,(1-\eta)r) \neq \emptyset$ if $\|x - \bar{x}(t)\| \leq \varepsilon$

for some $r > 0$, $R > 0$, $\eta > 0$, $\varepsilon > 0$ (depending on u). We further define:

$$\Delta(x(\cdot)) = \{t : g(t, x(t)) = \max_t g(t, x(t))\}$$

and

$\partial_C^{>} g(t,x) = \text{conv}\{\limsup \partial g(t_m, x_m) : t_m \to t,\ x_m \to x,\ g(t_m, x_m) > g(t,x)\}.$

We now have what we need to state the theorem.

Theorem 2

Under the assumptions (A_3), (A_4), (A_5) and assuming $\bar{x}(t) \in W^{1,1}$ is a local solution to the optimal control problem in Eq (3.1). Then there are $\lambda \geq 0$, a nonnegative measure μ on $[0,T]$ supported on $\Delta(\bar{x}(\cdot))$, a measurable $\mathbb{R}^n$-valued function $q(t)$ satisfying $\|q(t)\| < R(t)$ a.e., an $\mathbb{R}^n$-valued function $p(t)$ of bounded variation, continuous from the left, and a μ-measurable selection $\gamma(\cdot)$ of the set-valued mapping $t \to \partial_C^{>}(t, x(t))$ such that $dp(t) = q(t)dt + \gamma(t)\mu(dt)$ and the following four conditions are satisfied:

- (i) $\lambda + \|p(\cdot)\| + \mu([0,T]) = 1$ (*non-triviality*).
- (ii) $(p(0), -p(T) - \gamma(T)\mu(\{T\})) \in \lambda\partial l(\bar{x}(0), \bar{x}(T)) + N(S, (x(0), x(T)))$ (*transversality*).
  Here $N(S,(x(0),x(T)))$ denotes the (basic, limiting Mordukhovich) normal cone.
- (iii) $q(t) \in \text{conv}\{q : (q, p(t)) \in N(\text{Graph } F(t,\cdot), (\bar{x}(t), \dot{\bar{x}}(t)))\}$ a.e. on [0,T] (*Euler-Lagrange inclusion*)
- (iv) $\langle p(t), y - \dot{\bar{x}}(t)\rangle \leq 0, \forall y \in U(t)$ a.e. on [0,T] (*maximum principle*)

4 Main takeaway and further questions

The main takeaway is that a heavily constrained optimal control problem can be equivalently reduced to an unconstrained Bolza problem with simply structured integrand and off-integral term. As the author of this essay is mainly interested optimal control problems in the context of PDE-constrained optimization, one would like to extend the results to this framework. In particular it would be interesting to see if it was possible to prove that a heavily constrained optimal control PDE-constrained optimization problem could be equivalently reduced to a unconstrained PDE-constrained optimization and if it could, what conditions would apply.

One conjecture is that a constrained PDE-constrained optimization problem subject to state and/or control constraints can always be reformulated as a unconstrained PDE-OPT problem where the objective functional is necessarily non-differentiable.

References

Radon measure. *Encyclopedia of Mathematics*. URL http://www.encyclopediaofmath.org/index.php?title=Radon_measure.

S. Boyd and L. Vandenberghe. Notes on subgradients, 2008. URL https://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf.

A. D. Ioffe. On generalized bolza problem and its application to dynamic optimization. *Journal of Optimization Theory and Applications*, 182:285–309, 2019.

D. Klatte and B. Kummer. Aubin property and uniqueness of solutions in cone constrained optimization. *Mathematical Methods of Operational Research*, 77(3):291–304, 2013.

B. Messaoud. *Regularity Concepts in Nonsmooth Analysis*. Springer Optimization and Its Applications., 2012.

Bayesian Inversion of Dahlquist's test equation

Robin Eriksson*

September 3, 2019

Abstract

The Bayesian approach to inverse problems is a viable solution for some set-ups. A vital property of the Bayesian solution is that the resulting estimate includes a measure of uncertainty, which the classical approach omits. In this essay, we consider noisy observations from Dahlquist's test equation $z' = \lambda z$ and solve the inversion with the Bayesian approach, recovering an estimate for the posterior distribution of λ.

1 Introduction

The importance of having uncertainty quantification (UQ) in mind when developing numerical methods is sometimes wrongfully overlooked, as showed in Owhadi et al. [2015] where UQ and the data defines a threshold on the achievable numerical resolution. In short, developing UQ goes hand in hand with developing numerical solvers for actual data applications.

This essay covers the theoretical approach of Bayesian inversion, heavily inspired by Stuart and Taeb [2018], with an application on a toy model. In the model, we assume Gaussian measures for both the likelihood and the prior. These measures are conjugate, which leads to straightforward calculations of the posterior, further detailed in Section 2.

For non-conjugate measures, one has to employ computational techniques and often settle for approximations of the full distribution or finite samples, which one should take into consideration when solving for complex model and observation dynamics, see Reich and Cotter [2015] for examples.

*Division of Scientific Computing, Department of Information Technology, Uppsala university, SE-751 05 Uppsala, Sweden. robin.eriksson@it.uu.se

2 Bayesian Inversion

The Bayesian approach to inversion we follow leads to the notion of finding a probability measure containing information about the relative probability of different states $u \in X$, given the data $y \in Y$, where X and Y are separable Banach spaces. The probability measure is referred to as the *posterior measure* and denoted as $\mu^y(u)$, where $\rho(u|y)$ is its density.

The framework assumes that the data y are observations subject to noise defined as

$$y = \mathcal{G}(u) + \eta, \tag{2.1}$$

where η is the *observation noise* independent of u and $\mathcal{G}(\cdot)$ is the *observation operator*. We assume that η is a random variable with the density ρ_0. Then the probability of y given u has the density

$$\rho(y|u) := \rho(y - \mathcal{G}(u)). \tag{2.2}$$

One often refers to this density as the *data likelihood*. We describe our prior beliefs about u, in terms of the probability measure μ_0 and to its density $\rho(u)$.

In a probability setting, given a probability triplet $(\Omega, \mathcal{F}, \mathbb{P})$ where $A, B \in \mathcal{F}$ with $\mathbb{P}(A) > 0$, $\mathbb{P}(B) > 0$, one defines the conditional probability through Bayes' formula as

$$\mathbb{P}(A|B) = \frac{1}{\mathbb{P}(B)} \mathbb{P}(B|A)\, \mathbb{P}(A), \tag{2.3}$$

Suppose, if $(u, y) \in \mathbb{R}^d \times \mathbb{R}^\ell$ is a jointly distributed pair of random variables with Lebesgue density $\rho(u, y)$, then the infinitesimal version of (2.3) formula tells us that

$$\rho(u|y) \propto \rho(y|u)\rho(u), \tag{2.4}$$

where the normalization constant depends only on y. Thus

$$\rho(u|y) = \frac{\rho(y|u)\rho(u)}{\int_{\mathbb{R}^d} \rho(y|u)\rho(u)du}. \tag{2.5}$$

Abstractly, equation (2.4) express the relation between the posterior measure $\mu^y(u)$ (with density $\rho(u|y)$) and the prior measure $\mu_0(u)$ (with density $\rho(u)$) the Radon-Nikodym derivative [Stuart and Taeb, 2018, Thm. 6.2]

$$\frac{d\mu^y}{d\mu_0}(u) \propto \rho(y - \mathcal{G}(u)). \tag{2.6}$$

Additionally, since $\rho(\cdot)$ is a density and thus non-negative, without loss of abstraction we may write the right-hand side as the exponential of the negative of a potential $\Phi : X \times Y \to \mathbb{R}$, to obtain

$$\frac{d\mu^y}{d\mu_0}(u) \propto \exp(-\Phi(u; y)). \tag{2.7}$$

We impose some assumptions on the potential to formalize some essential properties of the posterior probability measure, namely it being well-defined and continuous concerning data. The assumptions follow.

Assumption 2.1. The function $\Phi : X \times Y \to \mathbb{R}$ has the following properties.

1. For every $\varepsilon > 0$ and $r > 0$ there is an $M = M(\varepsilon, r) \in \mathbb{R}$ such that, for all $u \in X$ and all $y \in Y$ with $||y||_Y < r$,
$$\Phi(u; y) \leq M - \varepsilon ||u||_X^2.$$
2. For every $r > 0$ there is a $K = K(r) > 0$ such that, for all $u \in X$ and $y \in Y$ with $\max\{||u||_X, ||y||_Y\} < r$,
$$\Phi(u; y) \leq K.$$
3. **(well-defined)** For every $r > 0$ there is an $L(r) > 0$ such that, for all $u_1, u_2 \in X$ and $y \in Y$ with $\max\{||u_1||_X, ||u_2||_X, ||y||_Y\} < r$,
$$|\Phi(u_1; y) - \Phi(u_2; y)| \leq L ||u_1 - u_2||_X.$$
4. **(continuity w.r.t. data)** For every $\varepsilon > 0$ and $r > 0$ there is a $C = C(\varepsilon, r) \in \mathbb{R}$ such that, for all $y_1, y_2 \in Y$ with $\max\{||y_1||_Y, ||y_2||_Y\} < r$, and for all $u \in X$,
$$|\Phi(u; y_1) - \Phi(u; y_2)| \leq \exp(\varepsilon ||u||_X^2 + C) ||y_1 - y_2||_Y.$$

Now consider the case where $\mathcal{G}$ is linear and that both the prior and the noise are Gaussian measures. We denote a Gaussian measure as $\mathcal{N}(m, C)$ with mean function m, and the covariance operator C. By employing (2.4), we see that the posterior is proportional to the product of two Gaussian measures, i.e., also a Gaussian measure, leading us to formulate the following theorem,

Theorem 2.1. *If the observation is linear, both the prior and the noise are Gaussian measures, and the noise is additive, then the posterior is also a Gaussian measure.*

One can determine the characterizing functions of the Gaussian measure that describe the posterior directly from [Stuart and Taeb, 2018, Thm. 6.20]. In summary, they render the conditional Gaussian distribution of x_1 given x_2 with the mean

$$m' = m_1 + \mathcal{C}_{12}\mathcal{C}_{22}^{-1}(x_2 - m_2), \tag{2.8}$$

and the covariance operator

$$\mathcal{C}' = \mathcal{C}_{11} - \mathcal{C}_{12}\mathcal{C}_{22}^{-1}\mathcal{C}_{21}, \tag{2.9}$$

where (x_1, x_2) is a Gaussian random variable with the mean $m = (m_1, m_2)$ and the covariance operator $\mathcal{C}$.

3 On Dahlquist's test equation

Consider the model dynamics defined by Dahlquist's test equation,

$$z' = \lambda z, \quad z(0) = 1. \tag{3.1}$$

We add an observation model on top of the model dynamics, as

$$y_i = \mathcal{G}_i(\lambda) + \eta_i, \quad \eta_i \sim \mathcal{N}(0, \gamma^2), \tag{3.2}$$

where $\mathcal{G}_i(\lambda) = z_i$, and $z_i = T_\lambda z_{i-1}$ for an operator T_λ. Since z is uniquely defined by λ, we can map λ to y and perform the inversion of the parameter through the observations. We are to compute the posterior density $\rho(\lambda|y_{1:N})$ given the set of noisy observations $y_{1:N} = \{y_{t_j}\}_{j=1}^N$ for $0 = t_1 < t_2 < \ldots t_N = 1$, where $y \in Y = \mathbb{R}^n$, and $\lambda \in X = \mathbb{R}$. We employ a Gaussian prior on λ with mean λ_0 and variance γ_0^2.

For this set-up, we see that both the prior and the noise are Gaussian measures, and the noise is additive, thus by Thm. 2.1 the posterior is also a Gaussian measure identified by mean m' and variance σ'^2,

$$m' = \left(\frac{\lambda_0}{\gamma_0^2} + \frac{\sum_{i=1}^N (y_i - \tilde{z}_i)}{\gamma^2}\right)\sigma'^{-2} \tag{3.3}$$

$$\sigma'^2 = \left(\gamma_0^{-2} + \gamma^{-2}\right)^{-1} \tag{3.4}$$

On the operator T_λ, it can be either exact or approximate. For the latter, one can denote it as $T_\lambda^{(h)}$ where h is the step length in the computational method, e.g., the Euler method, and $T_\lambda^{(h)} \to T_\lambda$ as $h \to 0$. How the approximation of T_λ, and implicitly the potential, affects the convergence of the posterior measure is further discussed in Stuart and Taeb [2018].

3.1 Error convergence

Consider an arrangement where we can increase the number of data points N. To observe error convergence with data, we perform the inversion and extract the maximum a posteriori (MAP) estimate, λ_{MAP}, and compute the error as the L_2 distance from the truth, λ_{true}. In Figure 3.1 A, we present the error together with a $N^{-0.5}$ line for reference, and see that they behave closely.

As described in the previous section, by the Bayesian approach we compute a posterior density $\rho(\lambda|y_{1:N})$. Using the posterior for $N = 100$, we give, in Figure 3.1 B, the 99% credible interval (CI) for z which from inspection is fairly decisive for the used noise variance.

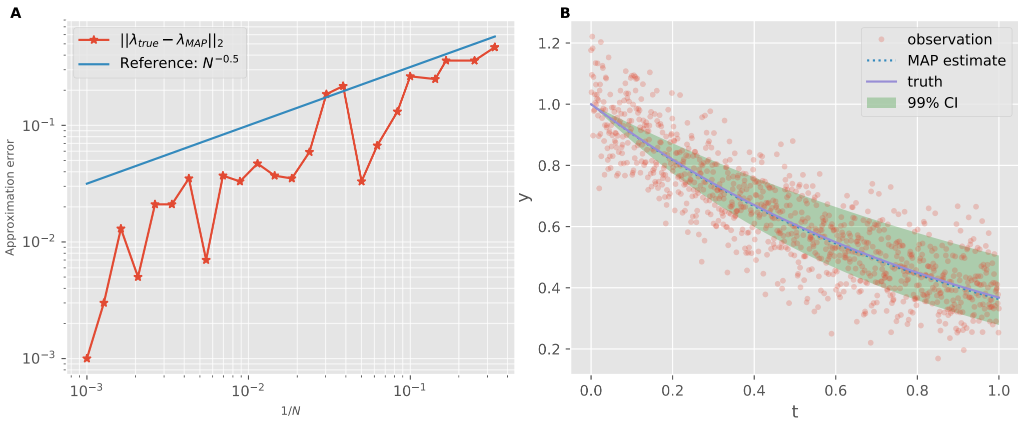


Figure 3.1: Two experiments with noise $\mathcal{N}(0, 0.1)$ and prior $\mathcal{N}(-2, 1)$. (A) MAP error against increasing number of observations N. (B) Data observation ($N = 100$) with MAP estimate and 99% Bayesian CI.

References

H. Owhadi, C. Scovel, and T. Sullivan. On the brittleness of bayesian inference. *SIAM Review*, 57(4):566–582, 2015.

S. Reich and C. Cotter. *Probabilistic forecasting and Bayesian data assimilation*. Cambridge University Press, 2015.

A. Stuart and A. Taeb. Data Assimilation and Inverse Problems. *arXiv:1810.06191v2*, 2018.

Condition number for compact operator equations

Fredrik Fryklund*

September 5, 2019

Abstract

We show that the eigenvalues for a compact linear operator can always be arranged as a sequence converging to zero. Such operators may arise in solving boundary integral equations and this property forms the foundation for demonstrating that the condition number of the discretised problem is bounded.

1 Introduction

Many partial differential equations can be formulated as a boundary integral equation, which in turn can be seen as operator equations $\alpha x - Tx = y$, where $T : X \to X$ is a compact linear operator on a normed space X. Here I is the identity operator, $y \in X$ and $\alpha \neq 0$ are given. By using a boundary integral method to approximate the operator and the solution, the equation is reduced to a linear system

$$Ax = y \tag{1.1}$$

to be solved for x. The condition number is a measure of how sensitive the solution is to small pertubations in the data. To keep things simple, we restrict ourself to partial differential equations where A is symmetric and real, such as exterior Laplace equation. Then the condition number $\mathrm{cond}(A) = |\lambda_{\max}|/|\lambda_{\min}|$, where λ denotes an eigenvalue of A.

Three different type of condition numbers can be distinguished for finite-dimensional approximations of operator equations. First we have the condition number of the original operator $\alpha I - T$, then of the approximating

*Department of Mathematics, KTH Royal Institute of Technology, Stockholm, Sweden. ffry@kth.se

operator and finally the finite-dimensional system $Ax = y$ that is solved numerically. In this essay we study the second one and the third one by relating them to the eigenvalues of T. As we shall see, the condition number is bounded and does not grow with successive refinement of the problem. This is a clear advantage over conventional methods such as the finite element method and the finite difference method.

2 On the accumulation of eigenvalues

Spectral theory is considerably more complicated for infinite dimensional-spaces X than for finite-dimensional spaces. The set of eigenvalues is now a subset of spectral values, which is a more complex structure to study. (We encourage the reader to refresh her knowledge on spectral theory, otherwise the full appreciation of the next statement is lost.) However, the spectrum of a compact operator is like that of a matrices, in the sense that it is countable and every nonzero spectral value is an eigenvalue Kreyszig [1989]. We will prove the former. Thus if $A_n x_n = y_n$ is a sequence of linear problems that in the limit goes to an operator equation, then the eigenvalues of A_n are approximations of the eigenvalues of the compact operator. We assume these approximations to be sufficiently accurate to bound the condition number of all A_n by considering the eigenvalues of the compact operator.

Theorem 2.1. *The set of eigenvalues of a compact linear operator $T : X \to X$ on a normed space X is countable (perhaps finite or even empty), and the only possible point of accumulation is $\lambda = 0$.*

Proof. A finite set is countable, thus we show that for each real $k_0 > 0$ there exists only a finite number of eigenvalues λ with $|\lambda| > k_0$. We do this by assuming the contrary, i.e. that there exists some $k_0 > 0$ such that the set of all $\lambda \in \sigma_p$, with $|\lambda| > k_0$, is infinite. By this assumption there exists an infinite sequence of eigenvalues (λ_n) such that $|\lambda_n| > k_0$ for all n. Each element λ_n of this sequence satsifies $Tx_n = \lambda_n x_n$, for some $x_n \neq 0$. Theorem $7.4 - 3$ from Kreyszig [1989] states that eigenvectors $x_1, \ldots, x_n$ corresponding to distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ of a linear operator T on a vector space X constitute a linearly independent set. This allows us to define finite-dimensional subspaces

$$M_n = \operatorname{span}(x_1, \ldots, x_n) \tag{2.1}$$

where each element $x \in M_n$ has a unique representation

$$x = \sum_{i=1}^{n} a_i x_i. \tag{2.2}$$

We now show that the operator $(T - \lambda_n I)$ maps elements from M_n to M_{n-1}. Consider

$$(T - \lambda_n I)\, x = \sum_{i=1}^{n} a_i T x_i - \sum_{i=1}^{n} a_i \lambda_n x_i = \sum_{i=1}^{n-1} a_i \left(\lambda_i - \lambda_n\right) x_i \tag{2.3}$$

and note that the right hand side is depends on x_i, with $i = 1, \ldots, x_{n-1}$, but is independent of x_n. Obviously it is an element of M_{n-1}, thus $(T - \lambda_n I)\, x \in M_{n-1}$ for all $x \in M_{n-1}$. For future reference we note that M_{n-1} is a proper subspace of M_n and since it is a finite-dimensional subspace of a normed space it is closed.

To finish the proof we construct a bounded sequence y_n such that the sequence Ty_n, by mapping y_n under the map T, does not have a convergent subsequence. Then, by theorem $8.3-1$ in Kreyszig [1989], the operator T is not a compact operator, which clearly is a contradiction. By Theorem $2.4-3$ in Kreyszig [1989] we can construct said sequence (y_n) by defining

$$y_n \in M_n, \quad \|y_n\| = 1, \quad \|y_n - y\| \geq \frac{1}{2} \text{ for all } y \in M_{n-1}. \tag{2.4}$$

Note that $y_m \in M_n$ for $n \geq m$. Therefore, for $m < n$ we have

$$Ty_n - Ty_m = \lambda_n y_n - (\lambda_n y_n - Ty_n + Ty_m) = \lambda_n \left(y_n - y\right) \tag{2.5}$$

where $y = \lambda_n^{-1} \tilde{y}$ and $\tilde{y} = (\lambda_n y_n - Ty_n + Ty_m)$. We now show that $\tilde{y} \in M_{n-1}$. Clearly $y_m \in M_m \subset M_{n-1} = \operatorname{span}\left(x_1, \ldots, x_{n-1}\right)$ since $m \leq n-1$, thus

$$Ty_m = \sum_{i=1}^{m} a_i T x_i = \sum_{i=1}^{m} a_i \lambda_i x_i \in M_m \subset M_{n-1}. \tag{2.6}$$

It remains to show that $\lambda_n y_n - Ty_n \in M_{n-1}$. From above we have that $(T - \lambda_n I) : M_n \to M_{n-1}$, thus

$$\lambda_n y_n - Ty_n = -\left(T - \lambda_n I\right) y_n \in M_{n-1} \tag{2.7}$$

and we have that $\tilde{y} \in M_{n-1}$ and clearly $y \in M_{n-1}$. Finally it holds that

$$\|T_n - T_m\| = \|\lambda_n \left(y_n - y\right)\| \geq |\lambda_n| \frac{1}{2} \geq \frac{1}{2} k_0, \tag{2.8}$$

by construction of y, see (2.4). We arrive at the conclusion that we can't find a convergent subsequence of (Ty_n) since (2.8) states that the distance between two elements is always greater or equal to $\frac{1}{2}k_0 > 0$. This is a contradiction, since T is indeed a compact operator, thus there is a convergent subsequence since (y_n) is bounded. We arrive at the conclusion that there does not exist a $k_0 > 0$ such that there are infinitely many $\lambda \in \sigma_p$ that satisfies $|\lambda| > k_0$. Instead this set is finite and therefore countable. It implies that 0 is the only possible accumulation point the eigenvalues can always be arranged into a sequence that converges to zero.

□

3 Condition number

We now demonstrate the connection of the eigenvalues with the condition number by considering the exterior Laplace equation

$$\Delta u(\boldsymbol{x}) = 0, \quad \boldsymbol{x} \in \Omega, \tag{3.1}$$

$$u(\boldsymbol{x}) = g(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega, \tag{3.2}$$

$$u(\boldsymbol{x}) = u_\infty, \quad |\boldsymbol{x}| \to \infty. \tag{3.3}$$

Here $\Omega \subset \mathbb{R}^2$ is an unbounded domain whose boundary $\partial\Omega$ is simply connected. Also, $\partial\Omega$ is assumed to be bounded and in C^2. We let ν denote the unit normal directed into Ω.

The classical way to reduce this boundary value problem to a Fredholm integral equation of the second kind is to seek the solution in the form of a double-layer potential

$$u(\boldsymbol{x}) = \int_{\partial\Omega} \frac{\partial K(\boldsymbol{x}, \boldsymbol{y})}{\partial n(\boldsymbol{y})} \sigma(\boldsymbol{y}) \, \mathrm{d}s_{\boldsymbol{y}}, \quad \boldsymbol{x} \in \Omega, \tag{3.4}$$

with an unknown density $\sigma \in C(\partial\Omega)$. Here the *kernel*

$$K(\boldsymbol{x}, \boldsymbol{y}) = -\frac{1}{2\pi} \log |\boldsymbol{x} - \boldsymbol{y}|, \tag{3.5}$$

denotes the fundamental solution to the Laplace equation in $\mathbb{R}^2$. The unknown density σ is obtained by solving the boundary integral equation

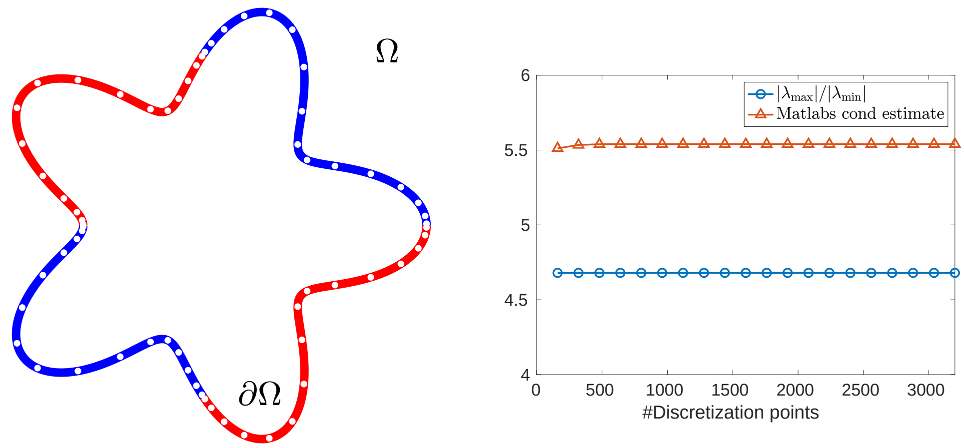
$$\frac{1}{2}\sigma(x) + \int_{\partial\Omega} \frac{\partial K(\boldsymbol{x}, \boldsymbol{y})}{\partial n_{\boldsymbol{y}}} \sigma(\boldsymbol{y}) \mathrm{d}s_{\boldsymbol{y}} = f(\boldsymbol{x}), \quad \boldsymbol{x} \in \partial\Omega, \tag{3.6}$$

or $(T - I)\sigma = -2f$, where $T\sigma$ is the second term in (3.6), scaled by -2, and is a linear compact operator. We claim without proof that there exists a unique solution σ to (3.6) Kreyszig [1989]. Similarily, the discretized problem reads

$$(A - I)x = y, \tag{3.7}$$

where A is a real and symmetric matrix. The eigenvalues of T can be arranged such that they converge to zero. Also, this result is needed for Theorem 10.21 in Kress [2014], that states that the eigenvalues of T lies in $[-1, 1)$. This could be problematic, since potentially $\lambda_{\min} = 0$, and the condition number becomes infinite. However, the eigenvalues of $(T - I)$ are shifted by one due to I. Thus for (3.6) the eigenvalues are in $[0, 2)$ and the condition number is bounded for all A_n. We stress that this holds for the Nyström method, while e.g. Galerkin methods may suffer from instabilities for a poor choice of basis. Furthermore, while the condition number does not grow with successive refinement of discretization of the operator equation it does depend on the geometry. The higher the curvature of the problem, the higher the condition number becomes.

4 Numerical experiments



To demonstrate the accumulation of eigenvalues and the boundedness of the condition number we solve (3.1)–(3.3) on a starfish shaped domain, see figure above, with the Dirichlet boundary data

$$g(\boldsymbol{x}) = \sum_{n=1}^{3} \frac{1}{|x - y_i|}. \tag{4.1}$$

Here y_i are randomly distributed points in the plane close to the boundary

$\partial\Omega$ but outside Ω. Each colored section of the starfish contains 16 Gauss-Legendre points, represented by the white dots. The given data and the solution are represented on these nodes. In the right image the condition number, as estimated by Matlab, and the computed condition number by $|\lambda_{\max}|/|\lambda_{\min}|$ are plotted against the number of Gauss-Legendre points. As predicted, the condition number goes to a constant value. Furthermore, $\lambda_{\max}$ is equal to 2, up to round off errors.

References

R. Kress. *Linear Integral Equations.* Applied Mathematical Sciences, 82. 3rd ed. 2014.. edition, 2014. ISBN 1-4614-9593-8.

E. Kreyszig. *Introductory functional analysis with applications.* Krieger, Malabar, Fla., repr. ed.. edition, 1989. ISBN 0-89464-332-0.

The universal approximation theorem

Fredrik Laurén *

August 22, 2019

Abstract

The universal approximation theorem is studied. The theorem states that one-layered neural networks can approximate any continuous function arbitrarily well. The Hahn-Banach Theorem and Riez's representation theorem is used for the proof, and both are hence introduced in a preliminary section. We focus on the part of the proof that is valid for general discriminatory functions. Then we state the result saying that sigmoidal functions are discriminatory. Finally, the proof for neural networks follows directly from the earlier stated results.

1 Introduction

Artificial neural networks have grown popular over the last few years because of impressive results where they are applied. However, it is challenging to understand how a neural network really works. If a network is trained well, we get high-quality results. On the other hand, if a network fails, it is hard to understand why it is doing so.

A feedforward neural network of an $n-$dimensional variable $\mathbf{x} \in \mathbb{R}^n$ is on the form (see for example Goodfellow et al. [2016])

$$y(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) \,. \tag{1.1}$$

In (1.1), $\alpha_i \in \mathbb{R}$, $\mathbf{w} \in \mathbb{R}^n$ are weights and $b_i \in \mathbb{R}$ are biases. Further,

$$\sigma(t) \to \begin{cases} 1 & \text{as} \quad t \to \infty \\ 0 & \text{as} \quad t \to -\infty \end{cases} \tag{1.2}$$

*Computational Mathematics, Department of Mathematics, Linköping University, SE-581 83 Linköping, Sweden. fredrik.lauren@liu.se

is a so-called sigmoidal function.

Let $C(I_n)$ denote the space of all continuous functions defined on the $n-$dimensional unit cube $I_n = [0,1]^n$. The aim of this mini essay is to study a theorem from Cybenko [1989] saying that sums of the form (1.1) are dense in $C(I_n)$ with respect to the supremum norm.

2 Preliminaries

For completeness, we include a few necessary theorems.

2.1 Hahn-Banach theorem

The theorem extends a function f from a subspace, W, to a function $\tilde{f}$, defined on the whole space, V. However, the focus here is on a version of the theorem that is used to separate two sets in a normed linear space, (see Rudin [1991]).

Theorem 2.1 (Hahn-Banach theorem (separation theorem))**.** *Let A and B be two non-empty and disjoint convex sets of a real normed linear space V. Assume A is closed and that B is compact. Then A and B can be separated strictly by a continuous linear functional f, i.e.*

$$f(x) \leq \alpha - \epsilon \quad \textit{and} \quad f(y) \geq \alpha + \epsilon \quad \forall x \in A\ ,\, \forall y \in B\,,$$

where $\alpha, \epsilon \in \mathbb{R}$ and $\epsilon > 0$.

From Theorem 2.1, the following proposition is deduced.

Proposition 2.2. *Let W be a subspace of a normed linear space V. Assume that $\overline{W} \neq V$. Then there exists a continuous linear functional f such that $f(x) = 0$ for all $x \in W$, but $f(y) \neq 0$ for all $y \in V$.*

Proof. Let $x_0 \in V \backslash \overline{W}$. Using Theorem 2.1, we set $A = \overline{W}$ and $B = x_0$. Since A is a subset, we have

$$tx + (1-t)y \in \overline{W} \quad \text{for all} \quad x, y \in \overline{W} \quad \text{and} \quad t \in [0,1]\,.$$

Hence, A is closed and B is compact and they are both non-empty disjoint convex sets. Thus, there exists a continuous linear functional f and $\alpha \in \mathbb{R}$ such that for all $x \in \overline{W}$

$$f(x) < \alpha < f(x_0)\,.$$

Since W is a linear subspace, $\lambda f(x) \leq \alpha$ for all $\lambda \in \mathbb{R}$ and $x \in W$ and hence $\alpha > 0$. However, if $\lambda = n$, where $n \in \mathbb{N}$, then

$$f(x) < \frac{\alpha}{n} \quad \forall\, x \in W \quad \Rightarrow f(x) \leq 0\,.$$

Further, $-x \in W$, thus $f(-x) = -f(x) \leq 0$. Therefore, $f(x) = 0$. Finally, we get $f(x) = 0 < \alpha < f(x_0)$, which concludes the proof. □

Proposition 2.2 provides a tool to examine if a subspace of a linear space is dense. Let V be a normed linear space, W a subspace of V and $x_0 \in V$. Then

$$x_0 \in \overline{W} \iff \begin{cases} \nexists \text{ continuous linear functional } f \text{ s.t.} \\ f(x) = 0 \quad \forall x \in W \quad \text{and} \quad f(x_0) \neq 0. \end{cases}$$

2.2 Riesz's Theorem

We focus on the version of the theorem defined on the space $C(I_n)$. The theory for this section can be found in Kreyszig [1978].

Definition 2.1. A function w defined on I_n is said to be of **bounded variation** on I_n if its *total variation* $\mathrm{Var}(w)$ on I_n is finite, where

$$Var(w) = \sup \sum_{j=1}^{n} |w(t_j) - w(t_{j-1})|\,.$$

The supremum is taken over all partitions

$$0 = t_0 \leq t_1 \leq \cdots \leq t_n = 1 \tag{2.1}$$

on the interval $[0, 1]$. Further, $n \in \mathbb{N}$ is arbitrary and so is the choice of values $t_1, \ldots, t_{n-1}$, as long as (2.1) is satisfied. All functions of bounded variation forms a vector space, $BV(I_n)$, with norm given by $\|w\| = |w(0)| + \mathrm{Var}(\omega)$.

Theorem 2.3 (Riesz's theorem - Functionals on $C(I_n)$)**.** *Every bounded linear functional f on $C(I_n)$ can be represented by the integral*

$$f(x) = \int_{I_n} x(t)dw(t)\,,$$

where w is of bounded variation on I_n and has the total variation

$$Var(w) = \|f\|.$$

3 Universal approximation theorem

In this section, the proof to the theorem from Cybenko [1989] is in focus.

Definition 3.1. We say that σ is *discriminatory* if for any $\omega \in BV(I_n)$

$$\int_{I_n} \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) d\omega(\mathbf{x}) = 0$$

for all $\mathbf{w}_i, \mathbf{x} \in \mathbb{R}^n, b_i \in \mathbb{R}$ implies that $\omega = 0$.

This definition tells us that for nonzero ω, there exist $\mathbf{w}_i$ and b_i such that

$$\int_{I_n} \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i) d\omega(\mathbf{x}) \neq 0\,.$$

We are now ready to state the Universal approximation theorem.

Theorem 3.1. *Let σ be any continuous discriminatory function. Then finite sums of the form (1.1) are dense in $C(I_n)$.*

Proof. We want to prove that the closure of the set of functions on the form (1.1) is all of $C(I_n)$.

Let $S \in C(I_n)$ be the set of functions of the form in (1.1). We see that the set S is a linear subspace of $C(I_n)$. The claim is that $\overline{S}$ is $C(I_n)$.

Assume the opposite, i.e. that $\overline{S}$, is not all of $C(I_n)$. Then, from Proposition 2.2, for some $h_0 \in C(I_n)$ there exists a continuous linear functional f such that

$$f(h_0) \neq 0 \quad \text{and} \quad f(h) = 0 \quad \forall h \in S\,. \tag{3.1}$$

From Theorem 2.3, we know that

$$f(h) = \int_{I_n} h(\mathbf{x}) d\omega(\mathbf{x})\,. \tag{3.2}$$

Combining (3.1) and (3.2) yields

$$f(h) = \int_{I_n} h(\mathbf{x}) d\omega(\mathbf{x}) = 0 \quad \forall h \in S\,. \tag{3.3}$$

Further, since $h = \sigma(\mathbf{w}^\top \mathbf{x} + b) \in S$, we get

$$f(\sigma(\mathbf{w}^\top \mathbf{x} + b)) = \int_{I_n} \sigma(\mathbf{w}^\top \mathbf{x} + b) d\omega(\mathbf{x}) = 0\,.$$

Since σ is discriminatory, $\omega(\mathbf{x}) = 0$. Hence,

$$f(h) = 0 \quad \forall x \in C(I_n)\,,$$

which is a contradiction. Thus, $\overline{S} = C(I_n)$. $\square$

We still need to know that the sigmoidal function σ in (1.1) is discriminatory. The proof for the following lemma can be found in Cybenko [1989].

Lemma 3.2. *Any bounded measurable sigmoidal function σ is discriminatory.*

Hence, Theorem 3.1 and Lemma 3.2 yield the following result.

Theorem 3.3. *Let σ be any continuous sigmoidal function. Then, finite sums of the form*

$$y(\mathbf{x}) = \sum_{i=1}^{N} \alpha_i \sigma(\mathbf{w}_i^\top \mathbf{x} + b_i)$$

are dense in $C(I_n)$. In other words, for any function $g \in C(I_n)$ and any $\epsilon > 0$, there is a sum $y(\mathbf{x})$ of the above form, for which

$$|y(\mathbf{x}) - g(\mathbf{x})| < \epsilon \quad \text{for all} \quad x \in I_n\,.$$

4 Consequences of the theorem

Theorem 3.3 states that a one-layered neural network can approximate any continuous function arbitrarily well. This is a powerful statement since a lot of problems can be reduced to functions. However, we did not mention anything about the number of terms in the sum. For many problems, using a one-layered network may lead to infeasibly large networks. Instead of using a single layer, several layers of smaller size can be used, which might increase the generalization ability of the network (see Goodfellow et al. [2016]).

References

G. Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

E. Kreyszig. *Introductory functional analysis with applications*, volume 1. wiley New York, 1978.

W. Rudin. *Functional Analysis*, volume 2. International series in pure and applied mathematics, 1991.

Adjoint-based a posteriori error estimation

Vidar Stiernström *

August 20, 2019

Abstract

A posteriori error estimates are fundamental in many mesh adaption algorithms. A useful framework for obtaining a posteriori error estimates for general quantities of interest is based on the theory of adjoint operators. This essay provides a brief outline of the theory of such adjoint-based a posteriori error estimates.

1 Introduction

Adaptive algorithms are often times useful for increasing the efficiency of a numerical method, be it by reduced computational cost, or improved accuracy. A key element in many adaptive algorithms lies in obtaining accurate a posteriori error estimates, i.e error estimates based on an computed approximation. In many cases specific quantities of solution are of particular interest and methods concerned with controlling the error of these quantities of interest (QoI) are referred to as goal-oriented methods (Grätsch and Bathe [2005]). A flexible framework for both computing a QoI and estimating its error rests on the theory of dual spaces, adjoint operators and Green's functions (Estep [2004]). Goal-oriented error estimates constructed using this framework are referred to as adjoint-based error estimates following the terminology of Kast [2017]. This essay attempts to outline some of the relevant theory on which adjoint-based a posteriori error estimation rests. Starting from the definition of dual spaces, the adjoint operator is first defined for normed spaces, and then specialized for Hilbert spaces, for which a method for formally obtaining the adjoint operator is presented. Next, in order to study QoIs, the generalized Green's function is introduced. Finally, adjoint-based a posteriori error estimation in the context of a general approximate solution is presented.

*Division of Scientific Computing, Department of Information Technology, Uppsala university, SE-751 05 Uppsala, Sweden. vidar.stiernstrom@it.uu.se

2 The adjoint operator

We start off by defining the dual of a normed space and the adjoint of a bounded linear operator (Kreyszig [1989]).

Definition 2.1. Let X be a normed space. The dual space of X, denoted X', is the set of all bounded linear functionals on X. X' is a normed space with norm defined by

$$\|f\|_{X'} = \sup_{\substack{x \in X \\ \|x\| \neq 0}} \frac{|f(x)|}{\|x\|} = \sup_{\substack{x \in X \\ \|x\| = 1}} |f(x)|. \tag{2.1}$$

Definition 2.2. Let $T : X \longrightarrow Y$ be a bounded linear operator, where X and Y are normed spaces. Then, for any bounded linear functional $g \in Y'$, the adjoint operator $T' : Y' \longrightarrow X'$ of T is defined by

$$f(x) = (T'g)(x) = g(Tx), \tag{2.2}$$

where X' and Y' are the dual spaces of X and Y, respectively.

The defining relation (2.2) is referred to as the *bilinear identity* (Estep [2004]). It follows from the Hahn-Banach theorem that T' is bounded with the norm $\|T'\| = \|T\|$ (Kreyszig [1989]). A short motivation for introducing the adjoint operator and its usefulness can be illustrated by the following example:

Let $T : X \longrightarrow Y$, be a bounded linear operator on normed vector spaces X and Y. Denote the range of T by $\mathcal{R}(T)$ and the nullspace of the adjoint operator T' by $\mathcal{N}(T')$. Consider the problem of finding $x \in X$ such that for $b \in Y$,

$$Tx = b.$$

Now, $\mathcal{R}(T)$ is the set of b for which there exists an solution, that is for $y \in \mathcal{R}(T)$, $\exists x$ such that $Tx = y$. It can be shown that a necessary condition for $y \in \mathcal{R}(T)$ is $y'(y) = 0 \quad \forall y' \in \mathcal{N}(T')$, i.e.

$$y \in \mathcal{R}(T) \Longrightarrow y'(y) = 0 \quad \forall y' \in \mathcal{N}(T').$$

Furthermore, if $\mathcal{R}(T)$ also is closed in Y, then the condition is sufficient, such that $y \in \mathcal{R}(T) \Longleftrightarrow y'(y) = 0 \quad \forall y' \in \mathcal{N}(T')$. Thus, T' provides information about the solvability of a problem (Estep [2004]).

The theory surrounding adjoint operators is indeed quite general, being defined on normed spaces. From here on we will restrict ourselves to the study of Hilbert spaces, which typically are of concern when considering solutions to partial differential equations. In this setting, the following representation theorem by Riesz is of importance (Kreyszig [1989])

Theorem 2.1. *Every bounded linear functional f on a Hilbert space H can be represented in terms of the inner product, namely*

$$f(x) = \langle x, z \rangle \tag{2.3}$$

where z depends on f, is uniquely determined by f and has a norm

$$\|z\| = \|f\|. \tag{2.4}$$

Now consider Definition 2.2, with $X = H_1$, $Y = H_2$, being Hilbert spaces. Then $T : H_1 \longrightarrow H_2$, and $T' : H_2' \longrightarrow H_1'$ with T' being defined by $T'g = f$, $g(Tx) = f(x)$, for $f \in H_1'$, $g \in H_2'$. In addition, by Theorem 2.1, the functionals f and g can be represented as $f(x) = \langle x, x_0 \rangle$, and $g(y) = \langle y, y_0 \rangle$, for some unique $x_0 \in H_1$, $y_0 \in H_2$. Via f, and g, now construct the operators

$$\begin{aligned} A_1 : H_1' \longrightarrow H_1, \quad &\text{defined by} \quad A_1 f = x_0 \\ A_2 : H_2' \longrightarrow H_2, \quad &\text{defined by} \quad A_2 g = y_0 \end{aligned}$$

One can show that A_1 and A_2 are bijective isometries (which follows directly from Theorem 2.1) and conjugate linear. This lets us define an operator $T^* : H_2 \longrightarrow H_1$ through the composition

$$T^* = A_1 T' A_2^{-1} \quad \text{defined by} \quad T^* y_0 = x_0$$

The operator T^* is referred to as the *Hilbert-adjoint*, and has the properties

$$\langle Tx, y_0 \rangle = \langle x, T^* y_0 \rangle \tag{2.5}$$

and $\|T^*\| = \|T\|$, such that T^* also is bounded. Indeed T^* can be defined directly through (2.5), and existence of T^* can be shown without going through the adjoint operator T' (Kreyszig [1989]).

Remark. In Estep [2004], Kast [2017] T^* is referred to simply as the adjoint. From here on this terminology will be adopted since only Hilbert spaces will be considered. It should however be noted that while T' is defined on the dual of the range of T, T^* is defined directly on the range of T.

3 Computing the adjoint

So far nothing has been stated regarding how to formally compute the adjoint. A simple example is that in the n-dimensional Euclidean spaces $\mathbb{R}^n$, for a linear operator T with matrix representation A the adjoint operator T' is represented by the transpose matrix $A^\top$, with the bilinear identity given by $y^\top Ax = x^\top A^\top y$ (Kreyszig [1989]). For Hilbert spaces, (2.5) is key in obtaining what is referred to as the *formal adjoint*. It is defined as follows (Estep [2004]).

Definition 3.1. Let L be a differential operator and let $H^s(\Omega)$ be a Hilbert space with sufficient smoothness s. The formal adjoint L^* is the differential operator that satisfies

$$\langle Lu, v\rangle_{L^2} = \langle u, L^* v\rangle_{L^2}, \quad \forall \text{ suitable } u, v \in H^s(\Omega). \tag{3.1}$$

Definition 3.1 is quite vague and a more precise definition can be given in terms of distribution theory, but this is out of scope for the current exposition. In short, the reason for requiring 'sufficient' smoothness and 'suitable' functions u, v is due to the fact that in general (3.1) can be satisfied even if T^* is not defined point-wise everywhere in Ω, and furthermore there are restrictions due to initial and/or boundary conditions in the setting of initial-boundary-value problems (Estep [2004]). It should be noted that relation (3.1) is also referred to as the bilinear (or adjoint) identity. The distinction from (2.2) will be clear from the context.

Now, using (3.1), the formal adjoint can be derived through integration by parts, which is best illustrated by an example. Consider

$$Lu = a(x)\tfrac{\partial u(x)}{\partial x} = f(x), \quad x \in \Omega \tag{3.2}$$

where u has compact support in Ω, and a is a real function. Taking the L^2 inner product with a test function v with compact support in Ω and integrating by parts results in

$$\langle Lu, v\rangle = \langle a\frac{\partial u}{\partial x}, v\rangle = auv\Big|_{\partial\Omega} - \langle u, \frac{\partial av}{\partial x}\rangle = \langle u, -\frac{\partial av}{\partial x}\rangle \tag{3.3}$$

Comparing (3.3) to (3.1) we immediately identify the formal adjoint as

$$L^* = -\frac{\partial}{\partial x}a.$$

For higher-order differential operators, integration by parts is carried out until the derivatives on u is moved to the test function v. For the case when u, v does not have compact support, v must be restricted to functions that are such that the bilinear identity still holds. For details, see Estep [2004], Kast [2017].

4 Adjoint-based a posteriori error estimation

We start off this section by introducing one final concept required for adjoint-based error estimation, namely the *generalized Green's function.* As the name suggests it is a generalization of the Green's function of a linear differential operator, which is tightly coupled to the formal adjoint, and in addition provides a means for studying QoIs via functionals. The generalized Green's function is defined as follows (Estep [2004]).

Definition 4.1. Let L be a linear differential operator and consider a problem of the form

$$\begin{aligned} Lu &= f, \quad x \in \Omega \\ u &= g, \quad x \in \partial\Omega \end{aligned} \tag{4.1}$$

where Ω is a space, time or space-time domain and $u = g$ are boundary/initial conditions such that the problem is well-posed. Let $\langle \cdot, \cdot \rangle$ denote the L^2 inner product. The generalized Green's function ϕ for a QoI q represented by $q = \langle u, \psi \rangle$ is

$$\begin{aligned} &L^* \phi(y, x) = \psi(x), \quad x \in \Omega \\ &\phi = g^*, \qquad\qquad\quad\; x \in \partial\Omega \end{aligned} \tag{4.2}$$

where L^* is the formal adjoint, and g^* are boundary/initial conditions such that $\langle Lu, \phi \rangle = \langle u, L^* \phi \rangle$ holds.

The motivation for introducing the generalized Green's function comes from the following relation

$$q = \langle u, \psi \rangle = \langle u, L^* \phi(y, x) \rangle = \langle Lu, \phi \rangle = \langle f, \phi \rangle, \tag{4.3}$$

where (4.2), and (4.1) was used together with the bilinear identity (3.1). Note that once the formal adjoint L^* is obtained and g^* is determined, (4.2) can be solved for ϕ independently of u and f for a given choice of ψ. Therefore (4.3) provides a means to compute q via ϕ. Furthermore, by considering different choices of ψ, different QoIs q can be observed. This observation serves as a basis for adjoint-based a posteriori error estimation, but is in general not limited to approximating errors.

If, for example, we are interested in how the solution at a particular interior point $x_0 \in \Omega$ respond to different forcings f, we choose $\psi = \delta_{x_0}$, since $u(x_0) = \langle u, \delta_{x_0} \rangle$, by the definition of the delta function. But then $u(x_0) = \langle \phi(x_0, x), f(x) \rangle$ by (4.3). Thus, once ϕ is obtained, it is possible to determine how $u(x_0)$ responds to different forcings simply by computing $\langle \phi, f \rangle$ for different f.

Having outlined the required theoretical tools, we now turn our attention towards adjoint-based a posteriori error estimation. Assume that for some numerical method, an approximate solution u_h to (4.1) is obtained. Then, an approximation of the QoI is computed by $q_h = \langle u_h, \psi \rangle$. Now consider the error in the estimated QoI which is given by $e_q = q - q_h$. Using (4.3), (4.2), the bilinear identity (3.1), and the sesquilinearity of the inner product, the following relation is derived:

$$\begin{aligned} e_q = q - q_h = \langle u, \psi \rangle - \langle u_h, \psi \rangle = \langle f, \phi \rangle - \langle u_h, L^* \phi \rangle \\ = \langle f, \phi \rangle - \langle L u_h, \phi \rangle = \langle f - L u_h, \phi \rangle \end{aligned} \tag{4.4}$$

Next, we introduce the residual r of (4.1). It is defined as $r(u_h) = Lu_h - f$ and is a measure of how well the solution u_h satisfies the partial differential equation. In particular, one can note that in the case $u_h = u$, then $r = 0$. However, since u_h is an approximation of u, r will be non-zero. Hence, by (4.4) the error in the QoI is given by

$$e_q = -\langle r(u_h), \phi \rangle, \tag{4.5}$$

which is referred to as the dual-weighted residual error estimate (Kast [2017], Grätsch and Bathe [2005]). Now, if we for instance would like to compute the error in the solution at an interior point x_0, we choose $\psi = \delta_{x_0}$ as above, solve (4.2) to obtain ϕ, and then compute the error by (4.5). Typically, the solution ϕ to (4.2) must also be computed via some numerical method. Then instead of an exact measure of the error e_q, an approximation of the error e_{q_h} is computed as $e_{q_h} = -\langle r(u_h), \phi_h \rangle$, with ϕ_h being the approximate solution to (4.2).

Remark. Other methods for estimating the error in a QoI can also be derived in the current framework, e.g energy norm based estimates. For an overview of methods for adjoint-based a posteriori error estimations used in FEM, see Grätsch and Bathe [2005].

References

D. J. Estep. A short course of duality, adjoint operators, green's functions, and a posteriori error analysis. *Lecture Notes*, 2004. URL https://pdfs.semanticscholar.org/a7f9/59110a2442e55b696d0b89c2d5cef5dcee51.pdf.

T. Grätsch and K.-J. Bathe. A posteriori error estimation techniques in practical finite element analysis. *Computers & Structures*, 83(4):235 – 265, 2005.

S. M. Kast. An introduction to adjoints and output error estimation in computational fluid dynamics. *Excerpt from PhD thesis*, 2017. URL http://arxiv.org/abs/1712.00693.

E. Kreyszig. *Introductory Functional Analysis with Applications*. Wiley Classics Library. Wiley, 1989.

Part II

Solutions to exercises

Chapter 1

Metric spaces

Problem 1.1.6

We consider the space l^∞, with the mapping $d(x,y) = \sup|x_i - y_i|$.
Show that $d(x,z) \leq d(x,y) + d(y,z)$ for any $x,y,z \in X$.

Given $x,y,z \in X$.
For each $i \in \mathbb{N}$, $|x_i - z_i| \leq |x_i - y_i| + |y_i - z_i|$.
So we deduce that $|x_i - z_i| \leq d(x,y) + d(y,z)$. This being true for any i, one gets that $d(x,z) \leq d(x,y) + d(y,z)$

Ex. 1.1.9

Show that d in 1.1-8 is a metric. (The discrete metric on $\mathbf{X}$ defined by $d(x,x) = 0, d(x,y) = 1, x \neq y$)

Solution

M1 $d(x,y) = 1 < \infty$. Therefore $d(x,y)$ is real valued, finite and non-negative.

M2 If $x = y, d(x,x) = 0$ follows from the definition of the discrete metric.
Conversely, if $d(x,y) = 0$, then $x = y$.

M3 $d(x,y) = d(y,x) = 1, x \neq y$.

M4 Let $z \in \mathbf{X}$,

$$1 \leq 1 + 1,$$
$$\therefore d(x,y) \leq d(x,z) + d(z,y).$$

Ex. 1.1.9

Show that d in 1.1-8 is a metric. (The discrete metric on $\mathbf{X}$ defined by $d(x,x) = 0, d(x,y) = 1, x \neq y$)

Solution

M1 $d(x,y) = 1 < \infty$. Therefore $d(x,y)$ is real valued, finite and non-negative.

M2 If $x = y, d(x,x) = 0$ follows from the definition of the discrete metric.
Conversely, if $d(x,y) = 0$, then $x = y$.

M3 $d(x,y) = d(y,x) = 1, x \neq y$.

M4 Let $z \in \mathbf{X}$,

$$1 \leq 1 + 1,$$
$$\therefore d(x,y) \leq d(x,z) + d(z,y).$$

Ex. 1.2.4

We are asked to find a sequence which converges to 0 but that is not in any space l^p, where $1 \leq p < \infty$. For this we recall that an element in the space l^p is a sequence of numbers such that:

$$\sum_{j=1}^{\infty} |\xi_j|^p < \infty$$

The immediate idea when having a divergent sum where the sequence goes to 0 is the harmonic sum, the addition of the powers makes this idea moot. What if we instead use:

$$\xi_j = \begin{cases} 0 \ ; \ j = 1 \\ \frac{1}{\log(j)} \ ; \ \text{else} \end{cases} .$$

Clearly we then have $\xi_j \to 0, \ (j \to \infty)$. We use the integral test

$$\int_2^{\infty} \frac{1}{\log(x)^p} dx =_{(*)} \int_{\log(2)}^{\infty} u^{-p} e^u du \to \infty$$

where $(*)$ marks the substitution $u = \log(x)$, $du = \frac{1}{x} dx$. We can conclude that the integral diverges as $\forall p \in [1, \infty)$:

$$\lim_{u \to \infty} u^{-p} e^u = \infty.$$

We thus have a sequence which converges to zero but is not in any l^p-space.

Ex. 1.2.11

If $(\mathbf{X}, d)$ is any metric space, show that another metric on $\mathbf{X}$ is defined by

$$\tilde{d}(x, y) = \frac{d(x, y)}{1 + d(x, y)},$$

and $\mathbf{X}$ is bounded in the metric d.

Solution

M1.

$$0 \leq \tilde{d}(x, y) = \frac{d(x, y)}{1 + d(x, y)} \leq 1 < \infty.$$

Therefore $\tilde{d}(x, y)$ is real valued, finite and non-negative.

M2. If $x = y$,

$$\tilde{d}(x, x) = \frac{d(x, x)}{1 + d(x, x)} = \frac{0}{1 + 0} = 0.$$

Conversely, if

$$\tilde{d}(x,y) = \frac{d(x,y)}{1+d(x,y)} = 0,$$

then $d(x,y) = 0$ which occurs when $x = y$ since $d(x,y)$ already defines a metric.

M3.

$$\tilde{d}(x,y) = \frac{d(x,y)}{1+d(x,y)} = \frac{d(y,x)}{1+d(y,x)} = \tilde{d}(y,x).$$

M4. Let $f(t) = \frac{t}{1+t}\ \forall\ t \in \mathbf{R}$. $f'(t) = (1+t)^{-2} = \frac{1}{(1+t)^2} > 0$ for $t > 0$. This shows that the function is monotone increasing.

Now,

$$\begin{aligned}
\tilde{d}(x,y) = \frac{d(x,y)}{1+d(x,y)} &\leq \frac{d(x,z)+d(z,y)}{1+d(x,z)+d(z,y)}, \\
&= \frac{d(x,z)}{1+d(x,z)+d(z,y)} + \frac{d(z,y)}{1+d(x,z)+d(z,y)}, \\
&\leq \frac{d(x,z)}{1+d(x,z)} + \frac{d(z,y)}{1+d(z,y)}, \\
&= \tilde{d}(x,z) + \tilde{d}(z,y).
\end{aligned}$$

Ex. 1.3.8

Show that the closure $\overline{B(x_0;r)}$ of an open ball $B(x_0,r)$ in a metric space can differ from the closed ball $\tilde{B}(x_0;r)$.

First we need to define the concepts (see Definition 1.3-1 in *Kreyszig*):

- Open ball: $B(x_0;r) = \{x \in X : d(x,x_0) < r)\}$,
- Closure of $B(x_0;r)$: $\overline{B(x_0;r)} = B(x_0;r) \cup \{$ all limit points to $B(x_0;r)\}$,
- Closed ball: $\tilde{B}(x_0;r) = \{x \in X : d(x,x_0) \leq r)\}$.

With the discrete metric

$$\hat{d}(x,y) = \begin{cases} 1 & x \neq x_0 \\ 0 & x = x_0 \end{cases}$$

and $r = 1$, we get

- $\overline{B(x_0;r)} = \underbrace{\{x \in X : d(x,x_0) < 1\}}_{x_0} \cup \underbrace{\{\text{ all limit points to } B(x_0;1)\}}_{x_0}$ - only the point x_0.
- $\tilde{B}(x_0;r) = \{x \in X : \hat{d}(x,x_0) \leq 1\} = X$ - the whole set X.

Thus, $\overline{B(x_0;r)}$ and $\tilde{B}(x_0;r)$ are not the same.

Ex. 1.3.8

Show that the closure $\overline{B(x_0;r)}$ of an open ball $B(x_0;r)$ in a metric space can differ from the closed ball $\tilde{B}(x_0;r)$.

Proof. Consider the discrete metric

$$d(x,x_0) = \begin{cases} 1, & x \neq x_0, \\ 0, & x = x_0, \end{cases}$$

over the set X. We have $B(x_0;1) = \{x_0\}$ and $\tilde{B}(x_0;1) = M$. Let x be an accumulation point of $B(x_0;1)$. Since x is an accumulation point, and $B(x_0;1) = \{x_0\}$, it means that $d(x,x_0) < \epsilon$ for any $\epsilon > 0$. However, for $\epsilon < 1$ the inequality $d(x,x_0) < 1$ is only satisfied for $x = x_0$. Thus $\overline{B(x_0;1)} = B(x_0;1)$, but $\overline{B(x_0;1)} \subset \tilde{B}(x_0;1) = X$. □

Ex. 1.3.12

Show that $B[a,b]$, where $a < b$, is not separable.

Solution. Consider

$$f_x = \begin{cases} 1, & x = x_0 \text{ where } x_o \in [a,b] \\ 0, & \text{otherwise.} \end{cases} \tag{1.0.1}$$

All $f_x(x)$ are in $B[a,b]$ and $d(f_x(x), f_y(y)) = 1$ when $x \neq y$.

All open balls $B(f_x(x);0.5)$ are disjoint, and, if D is dense, D has to intersect all these open balls in different points.

We have $|D| \geq |[a,b]|$. As the interval is uncountable, there cannot be a countal dense set, i.e., the space is not separable.

Ex. 1.3.12

Show that the space $B[a,b]$ with $d(x,y) = \sup|x(t) - y(t)|$ is not separable.

Let's define for any $x \in [a,b]$ the function $e_x = \delta_x$ the Kronecker function at x.
Then the family $(e_x)_{x\in[a,b]}$ is uncountable, and for $x \neq y$, $d(e_x, e_y) = 1$.
Thus the open balls of radius $\frac{1}{2}$ and of center e_x form an uncountable collection of disjoint open sets in $B[a,b]$.
Therefore, $B[a,b]$ is not separable.

Ex. 1.4.2

If (x_n) is Cauchy and has a convergent subsequence, say, $x_{n_k} \to x$ show that (x_n) is convergent with the limit x.

(x_n) is Cauchy if for any $\frac{\varepsilon}{2} > 0$ there is $N_1 = N_1(\varepsilon)$ such that:

$$d(x_m, x_n) < \frac{\varepsilon}{2} \quad \forall m, n > N_1 \tag{1.0.2}$$

We know that $x_{n_k} \to x$. Therefore $\forall \varepsilon > 0$ there exists $N_2 = N_2(\varepsilon)$:

$$d(x_{n_k}, x) < \frac{\varepsilon}{2} \quad n_k > N_2 \tag{1.0.3}$$

To show that $x_n \to x$ we need $\forall \varepsilon > 0$ there is an N such that

$$d(x_n, x) < \varepsilon, \quad n > N.$$

We have that:

$$\begin{aligned} d(x_n, x) &\leq d(x_{n_k}, x_n) + d(x_{n_k}, x) \\ &\leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon \end{aligned}$$

for $n_k > N_1$ and $n > N_2$. This proves the $d(x_n, x) < \varepsilon$ criteria for convergence of x_n to x.

Ex. 1.4.2

If (x_n) is Cauchy and has a convergent subsequence, say, $x_n \to x$, show that (x_n) is convergent with the limit x.

Solution. Take $\varepsilon > 0$, then there is K s.t.

$$|x_{n_k} - x| < \frac{\varepsilon}{2}, \text{ for all } k \geq K.$$

Moreover, since (x_n) is Cauchy, there is N s.t.,

$$|x_n - x_m| < \frac{\varepsilon}{2}, \text{ for all } m, n \geq N.$$

Take k s.t. $n_k \geq N$ and let $n \geq N$.

Then

$$|x_n - x| \leq |x_n - x_{n_k}| + |x_{n_k} - x| < \frac{\varepsilon}{2} + \frac{\varepsilon}{2}| = \varepsilon.$$

Hence (x_n) is convergent with the limit x.

Ex. 1.5.6

Show that the set of all real numbers constitutes an incomplete metric space if we choose

$$d(x, y) = |\arctan(x) - \arctan(y)|.$$

Solution

A metric space X is said to be complete if every Cauchy sequence in X converges. In the limit $x_n \to \infty$, $\arctan(x_n) \to \pi/2$, hence there exists a Cauchy sequence

$$|\arctan(x_m) - \arctan(x_n)| \leq \epsilon,$$

that converges as $x_n \to \infty$. Since ∞ is not in R, the suggested metric space is not complete.

Ex. 1.5.6

Show that the set of all real numbers constitutes an incomplete metric space if we choose:

$$d(x, y) = |\arctan(x) - \arctan(y)|$$

We recall that a metric space (X, d) is complete if every Cauchy-sequence converges, that is, has a limit which is an element of X. We thus need to find a Cauchy-sequence $(x)_n$ in $(\mathbb{R}, d)$ so that:

$$x_n \to x \notin \mathbb{R}$$

The obvious peculiarity with arctan is the limits:

$$\lim_{x\to\infty} \arctan(x) = \frac{\pi}{2}$$

so the sequence $(x)_n = n$ is Cauchy with respect to the given metric yet the sequence is divergent which is what we wanted to show. We can formalize it a bit better as follows: Given $\epsilon > 0$ choose $N > \arctan(\frac{\pi}{2} - \epsilon)$ so for $n, m > N$ we have:

$$d(x_n, x_m) = |\arctan(n) - \arctan(m)| \leq |\frac{\pi}{2} - \arctan(N)| < |\frac{\pi}{2} - \frac{\pi}{2} + \epsilon| = \epsilon$$

so the sequence is Cauchy yet divergent.

Note: Another way would be to use that $\arctan : \mathbb{R} \to (-\frac{\pi}{2}, \frac{\pi}{2})$ is a bijective map and to note that the metric space $((-\frac{\pi}{2}, \frac{\pi}{2}), d_I(x, y))$ where $d_I(x, y) = |x - y|$ is incomplete by standard arguments. (here we also use that this choice of metric gives us an isometry, i.e., a distance perserving transformation with regard to the metrics).

Ex. 1.5.8

Let be $Y = \{x \in C[a,b]\, s.t.\, x(a) = x(b)\}$ with the distance $d(x,y) = \sup |x(t) - y(t)|$. Show that Y, d is complete

We already know that $C[a,b]$ is complete with regard to this metric.
Then we just have to show that Y is a closed subset of $C[a,b]$. If (x_n) is a sequence of Y which converges to a limit $x \in X$, then we have:

$$x(a) = \lim x_n(a) = \lim x_n(b) = x(b)$$

and so $x \in Y$. So Y is complete

Ex. 1.6.6

Show that $C[0,1]$ and $C[a,b]$ are isometric.

Solution

Any function $\tilde{x}(t) \in C[a,b]$ can be mapped to a function $x(t) \in C[0,1]$ by $x(t) = T\tilde{x}(t) = \tilde{x}((t-a)/(b-a))$. Since T is a bijective mapping $C[1,0]$ and $C[a,b]$ are isometric.

Ex. 1.6.6

Show that $C[0,1]$ and $C[a,b]$ are isometric.

Solution: let us denote $C[0,1] = X_1$, $d_1(x,y) = \sup\limits_{t\in[0,1]} |x(t)-y(t)|$, $C[a,b] = X_2$, $d_2(\tilde{x},\tilde{y}) = \sup\limits_{t\in[a,b]} |\tilde{x}(t) - \tilde{y}(t)|$. (X_1, d_1) is a metric space, (X_2, d_2) is also a metric space. We need to find a bijective mapping $T : X_1 \longrightarrow X_2$ such that

$$d_2(Tx, Ty) = d_1(x,y). \tag{1.0.4}$$

Consider the following mapping from $[a,b]$ to $[0,1]$:

$$\psi(t) = (t-a)/(b-a). \tag{1.0.5}$$

Its inverse is $\psi^{-1}(x) = (b-a)x + a$. Mapping $T : X_1 \longrightarrow X_2$ can be defined as acting by mapping ψ^{-1} to the function's argument. The inverse mapping is then acting by ψ to the functions argument. Then

$$\begin{aligned}
d_2(Tx, Ty) &= \sup_{t\in[a,b]} |Tx(t) - Ty(t)| && (1.0.6)\\
&= \sup_{t\in[a,b]} |x((t-a)/(b-a)) - y((t-a)/(b-a))| && (1.0.7)\\
&= \sup_{t\in[0,1]} |x(t) - y(t)| = d_1(x,y). && (1.0.8)
\end{aligned}$$

To prove that T is bijective, we need to show that it is both injective and surjective.

Injectivity (maps distinct elements to distinct): if $x = y$ then $0 = d_1(x, y) = d_2(Tx, Ty)$, thus $Tx = Ty$.

Surjectivity (takes all possible values): for any $x \in X_2$ $T^{-1}x \in X_1$ and $T(T^{-1}x) = x$. $\square$

Ex. 1.6.14

Consider

$$d(x, y) = \int_a^b |x(t) - y(t)| \, \mathrm{d}t. \tag{1.0.9}$$

The question is if d defines a metric or pseudometric on a set X.

(i) Let X be the set of all real-valued continuous functions on $[a, b]$. Problem 1.1.8 shows that d defines a metric on X in this case.

(ii) Let X be the set of all real-valued Riemann integrable functions on $[a, b]$. A function $f : [a, b] \to \mathbb{R}$ is Riemann integrable (in the proper sense) if and only if it is bounded and the set of points where it is discontinuous has measure zero.

- If x and y are Riemann integrable then $|x(t) - y(t)|$ is also Riemann integrable, which means that $d(x, y)$ has a value in the real numbers. Since the integrand is nonnegative, $d(x, y) \in [0, \infty)$. This shows that (M1) is satisfied.
- It is clear that $d(x, x) = 0$. Conversely, if $d(x, y) = 0$ then $x = y$ almost everywhere (x and y are equal everywhere except possibly in a set of measure zero). Since x and y are allowed to have discontinuities it could be that there are isolated points where they differ. This means that $d(x, y) = 0$ does *not* imply $x = y$. Hence (M2) is *not* satisfied, but axiom (M2*) of a pseudometric is satisfied.
- Clearly $d(x, y) = d(y, x)$, so (M3) is satisfied.
- The proof of (M4) from Problem 1.1.8 goes through without modification in this setting.

This means that d defines a pseudometric on X in this case.

Ex. 1.6.14

Which kind of metric (regular or pseudo) is $d(x, y) = \int\limits_a^b |x(t) - y(t)| dt$ if

1. $X = C[a, b]$;
2. $X = \mathcal{R}[a, b]$, i.e. the space of Riemann-integrable functions on $[a, b]$.

Solution: The difference between a regular metric and a pseudo-metric is that the latter can attain zero on two distinct elements of the set. $d(x, y)$ is a regular metric if $d(x, y) = 0 \Leftrightarrow x = y$, and $d(x, y)$ is a pseudo-metric if $d(x, y) = 0 \Leftarrow x = y$.

1. Consider $X = C[a, b]$.

   First, we show "$\Leftarrow$": assume $x(t) = y(t)\ \forall t \in [a, b]$. Thus $|x(t) - y(t)| = 0\ \forall t \in [a, b]$ and $\int_a^b |x(t) - y(t)|dt = 0$, i.e. $d(x, y) = 0$.

   Second, we show "$\Rightarrow$": assume $d(x, y) = 0$, i.e. $\int_a^b |x(t) - y(t)|dt = 0$. Assuming that $a < b$, and taking into account that the integrand is continuous (this can be also shown), we conclude that $|x(t) - y(t)| = 0\ \forall t \in [a, b]$ and thus $x(t) = y(t)\ \forall t \in [a, b]$.

   This proves that $d(x, y)$ is a regular metric on $C[a, b]$.

2. Consider $X = \mathcal{R}[a, b]$.

   Showing "$\Leftarrow$" is similar to the previous case and therefore is omitted.

   To show that $d(x, y) = 0 \not\Rightarrow x = y$ we need to recall that a Riemann-integrable function is a bounded function continuous almost everywhere, i.e. with no more then countable number of discontinuities.

   It is enough to give one example to show that $d(x, y) = 0 \not\Rightarrow x = y$ in general. Let $a = 0, b = 1, y(t) = 0$ and
   $$x(t) = \begin{cases} 0, \text{if } t \neq 0.5, \\ 1, \text{if } t = 0.5. \end{cases}$$
   Both $x(t)$ and $y(t)$ belong to $\mathcal{R}[a, b]$, as well as $|x(t) - y(t)|$. In this case, $d(x, y) = \int_a^b |x(t) - y(t)|dt = 0$, whereas $x(t) \neq y(t)$.

   This proves that $d(x, y)$ is a pseudo-metric on $\mathcal{R}[a, b]$.

Problem 1.6.14

Let's consider $d(x, y) = \int_a^b |x(t) - y(t)|dt$. Show that d is a metric on $C[a, b]$, but only a pseudometric on the space of Riemann integrable functions on $[a, b]$.

- The positivity and symmetry of d is obvious. The triangle inequality for $|\cdot|$ gives naturally the triangle inequality for d.

- In the space of continuous functions, one can state that if $d(x, y) = 0$, then $x = y$ almost everywhere, and the continuity allows us to state $x = y$. So d is then a metric.

- In the space of Riemann-integrable functions, we do only have almost everywhere equality, which does not imply equality. Thus d is only a pseudometric.

Chapter 2

Normed spaces

Ex. 2.2.13

Show that the discrete metric on a vector space $\mathbf{X} \neq \{0\}$ cannot be obtained from a norm. (Cf. 1.1-8.)

Solution

The discrete metric is denoted

$$d(x,x) = 0, d(x,y) = 1, x \neq y.$$

For any $x, y \in \mathbf{X}, x \neq y$ and $\alpha \in \mathbf{R}$ we have $d(\alpha x, \alpha y) = 1, \alpha x \neq \alpha y$. However, $\alpha d(x,y) = \alpha \cdot 1 = \alpha, x \neq y$. Therefore the discrete metric cannot be obtained from a norm because it violates N3, that is;

$$d(\alpha x, \alpha y) = 1 \neq \alpha d(x,y) = \alpha.$$

Ex. 2.3.10

Show that if a normed space has a Schauder basis, it is separable.

Proof:
By definition, a space X is separable if it has a countable subset M, which is dense in X. Now we assume that X is a normed space in which there is a Schauder basis. Then X contains a sequence (e_n) such that

$$\forall x \in X \exists (a_n) : \lim_{n\to\infty} \|x - \sum_{i=1}^{n} a_i e_i\| = 0,$$

where the sequence (a_n) is unique. We then write $x = \sum_{i=1}^{\infty} a_i e_i$. Without restriction we can assume $\|e_n\| = 1 \quad \forall n$, otherwise we just re-normalize the basis. Furthermore, since the series is convergent, $\forall \varepsilon > 0, \quad \exists N > 0$ such that for $\forall n > N$, $x \in X$,

$$\|x - \sum_{i=1}^{n} a_i e_i\| < \frac{\varepsilon}{2}.$$

Now consider the subset

$$M = \{y \in X, b_i \in Q, n \in \mathbb{Z}^+ : y = \sum_{i=1}^{n} b_i e_i\}.$$

Depending on whether X is a real or complex normed space we choose $Q = \mathbb{Q}$ or $Q = \{a + ib : a, b \in \mathbb{Q}\}$, such that Q is dense in the field K of X. Then any $a_i \in K$ can be approximated arbitrarily well by an element in Q, i.e for $\forall a_i \in K \quad \exists b_i \in Q, \quad \varepsilon > 0$ such that

$$|a_i - b_i| < \frac{\varepsilon}{2^{i+1}}.$$

Therefore, for $x \in X$ and $y \in M$, by the triangle inequality

$$\begin{aligned}
\|x - y\| &\leq \|x - \sum_{i=1}^{n} a_i e_i\| + \|\sum_{i=1}^{n} a_i e_i - \sum_{i=1}^{n} b_i e_i\| \\
&\leq \|x - \sum_{i=1}^{n} a_i e_i\| + \sum_{i=1}^{n} |a_i - b_i| \|e_i\| \\
&< \frac{\varepsilon}{2} + \sum_{i=1}^{n} \frac{\varepsilon}{2^{i+1}} \\
&< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon,
\end{aligned}$$

where $\|e_i\| = 1$, and that the partial geometric series is bounded by the infinite series (which converges to $\varepsilon/2$), was used to obtain the two last inequalities. This shows that an element in X can be arbitrarily well approximated by an element in M, i.e M is dense in X. Moreover, M is countable since both $\{(e_n)\}$ and Q is. Thus X is by definition separable.

Ex. 2.4.4

Show that equivalent norms on a vector space X induce the same topology for X.

Proof. Let $\|\cdot\|_a$ and $\|\cdot\|_b$ be two different but equivalent norms, meaning that there exists positive numbers c_1 and c_2 such that for every x in X we have

$$c_1 \|x\|_a \leq \|x\|_b \leq c_2 \|x\|_a.$$

Moreover, let $(X, \mathcal{T}_a)$ and $(X, \mathcal{T}_b)$ be the topological spaces induced by $\|\cdot\|_a$ and $\|\cdot\|_b$, respectively.

Consider the open ball $B(x_0, r)_i = \{x \in X \mid \|x_0 - x\|_i < r\}$, $i = a, b$. For any $r > 0$ we have

$$\begin{aligned}
\|x\|_a \leq r &\Rightarrow \|x\|_b \leq c_2 r, \\
\|x\|_b \leq r &\Rightarrow \|x\|_a \leq \frac{r}{c_1},
\end{aligned}$$

that is

$$\begin{aligned}
B(x_0, r)_a &\subseteq B(x_0, c_2 r)_b, \\
B(x_0, r)_b &\subseteq B(x_0, r/c_1)_a.
\end{aligned}$$

An open subset is a union of balls, thus a collection $\mathcal{T}$ of all open subsets of X consists of all open balls in X. Since any ball in either norm can be scaled such that it is included in the other, there is no open ball in $\mathcal{T}_a$ that is not in $\mathcal{T}_b$, and vice versa. Thus the topologies $(X, \mathcal{T}_a)$ and $(X, \mathcal{T}_b)$ are the same. $\square$

Ex. 2.7.2

Show that a linear operator $T : X \longrightarrow Y$ is bounded $\Leftrightarrow$ it maps bounded sets in X to bounded sets in Y.

Solution: Let us first consider proof "$\Rightarrow$". T is bounded, thus

$$\exists c \in \mathbb{R}, c < +\infty : \|Tx\| \leq c\|x\|,\ \forall x \in X.$$

Let M be any bounded subset of X. It can be placed inside a ball of a finite radius with center at $0 \in X$, although 0 might not belong to M. Thus $\forall x \in M\ \ d(x, 0) = \|x - 0\| = \|x\| < +\infty$. Then

$$\|Tx\| \leq c\|x\| < +\infty\ \forall x \in M,$$

i.e. the image of M is bounded. Since M is an arbitrary bounded subset of X, one concludes that for any bounded subset of X the image is a bounded subset of Y.

Let us now prove "$\Leftarrow$." Let $M \subset X$ be bounded as well as $N \subset Y$, $T : M \longrightarrow N$. Then $\exists R_M, R_N$ such that $\forall x \in M\ \|x\| \leq R_M < +\infty$ and $\|Tx\| \leq R_N < +\infty$. Now consider any $\hat{x} \neq 0 \in M$ and $x = R_M \frac{\hat{x}}{\|\hat{x}\|}$, $\|x\| = R_M$. Then

$$\frac{R_M}{\|\hat{x}\|}\|T\hat{x}\| = \left\|T\left(R_M \frac{\hat{x}}{\|\hat{x}\|}\right)\right\| = \|Tx\| \leq R_N,$$

$$\|T\hat{x}\| \leq \frac{R_N}{R_M}\|\hat{x}\|.$$

This proves that T is bounded. $\square$

Ex. 2.7.7

(**Inverse operator**) Let T be a bounded linear operator from a normed space $\mathbf{X}$ onto a normed space $\mathbf{Y}$. If there is a positive b such that $\|Tx\| \geq b\|x\|$ for all $x \in \mathbf{X}$, show that then $T^{-1} : \mathbf{Y} \to \mathbf{X}$ exists and is bounded.

Solution

We show that T is injective, and therefore T^{-1} exists. T is already surjective since it is a linear operator. Choose any $x_1, x_2 \in \mathbf{X}, x_1 \neq x_2$. We have

$$\|Tx_1 - Tx_2\| = \|T(x_1 - x_2)\| \leq b\|x_1 - x_2\| > 0.$$

Since $x_1 \neq x_2$, this means $Tx_1 \neq Tx_2$.
To show that T^{-1} is bounded, we find $c > 0$ such that $\|T^{-1}y\| \leq c\|y\|$ for all $y \in \mathbf{Y}$. Since T is surjective, for any $y \in \mathbf{Y}$, there exists $x \in \mathbf{X}$ such that $y = Tx$ and if T^{-1} exists, then $x = T^{-1}y$. Therefore

$$\begin{aligned} \|T(T^{-1}(y))\| &\geq b\|T^{-1}y\|, \\ \|y\| &\geq b\|T^{-1}y\|, \\ \|T^{-1}y\| &\geq \frac{1}{b}\|y\| = c\|y\|, c = \frac{1}{b} > 0. \end{aligned} \tag{2.0.1}$$

Therefore T^{-1} is bounded.

Ex. 2.7.8

Show that the inverse $T^{-1} : \mathcal{R}(T) \to X$ of a bounded linear operator $T : X \to Y$ need not be bounded. Hint: Use T in Prob. 2.7.5.

The operator in Problem 2.7.5 is given by:

$$T : l^\infty \to l^\infty \quad y = (n_j) = Tx, \quad n_j = \xi_j/j, \quad x = \xi_j\,.$$

We first want to prove that T is bounded and linear:

$$\begin{aligned} \text{linear: } T(ax + by) &= T(a\xi_j + b\eta_j) = \frac{a\xi_j + b\eta_j}{j} = a\frac{\xi_j}{j} + b\frac{\eta_j}{j} = aTx + bTy, \\ \text{bounded: } \|Tx\| &= \sup_j \left|\frac{\xi_j}{j}\right| \leq \sup_j |\xi_j| = \|x\|\,. \end{aligned}$$

Further,

$$Tx = Ty \iff T(x - y) = 0 \iff x = y\,.$$

Hence, $Tx = 0 \Rightarrow x = 0$, so T^{-1} exists (Theorem 2.6-10).

Now, for any $n \in \mathbb{N}$, the sequence

$$(x_n) = \left(\frac{1}{1}, \frac{1}{\sqrt{2}}, \dots, \frac{1}{\sqrt{n}}, 0, 0, \dots\right) \in l^\infty$$

and we get

$$y_n = Tx_n = \left(1, \frac{1}{\sqrt{2}}, \dots, \frac{1}{\sqrt{n}}, 0, 0, \dots\right)\,.$$

We have $\|x\| = \sup_j x_j = \sqrt{n}$ and $\|y\| = \|Tx\| = \sup_j y_j = 1$. iSo,

$$\forall_{a>0\in\mathbb{R}},\ \exists_{n\in\mathbb{N}}:\ \sqrt{n} > a\,,$$

which yields

$$\|T^{-1}y_n\| = \|x_n\| = \sqrt{n} > c = c\|y\|\,,$$

so T^{-1} is not bounded.

Ex. 2.10.4

Let (T_n) be a sequence of bounded linear operators from a normed space $(X, \|\cdot\|_X)$ into a normed space $(Y, \|\cdot\|_Y)$. We shall show that if $T_n \to T$, then for every $\varepsilon > 0$ and any given closed ball B in X there is an N such that $\|T_n x - Tx\|_Y < \varepsilon$ for all $n > N$ and all $x \in B$.

$T_n \to T$ means that $\|T_n - T\|_{\text{op}} \to 0$ as $n \to \infty$, where

$$\|T\|_{\text{op}} = \sup_{\substack{x\in X\\ x\neq 0}} \frac{\|Tx\|_Y}{\|x\|_X}. \tag{2.0.2}$$

Thus, for any $\tilde{\varepsilon} > 0$ there is some $N_{\tilde{\varepsilon}}$ such that

$$\|T_n - T\|_{\text{op}} = \sup_{\substack{x\in X\\ x\neq 0}} \frac{\|T_n x - Tx\|_Y}{\|x\|_X} < \tilde{\varepsilon} \tag{2.0.3}$$

for all $n > N_{\tilde{\varepsilon}}$. This in turn implies that

$$\|T_n x - Tx\|_Y < \tilde{\varepsilon}\|x\|_X \tag{2.0.4}$$

for all $n > N_{\tilde{\varepsilon}}$ and all $x \neq 0$.

Now pick any $\varepsilon > 0$ and any closed ball $B \subset X$. Since B is bounded, there is some M such that $\|x\|_X \leq M$ for all $x \in B$. Take any $x \in B$. If $x = 0$, then $\|T_n x - Tx\|_Y = 0 < \varepsilon$ since T_n and T are linear. If $x \neq 0$, then (2.0.4) with $\tilde{\varepsilon} = \varepsilon/M$ implies that

$$\|T_n x - Tx\|_Y < \frac{\varepsilon}{M}\|x\|_X \leq \frac{\varepsilon}{M} M = \varepsilon \tag{2.0.5}$$

for all $n > N_{\varepsilon/M} = N$. This concludes the proof.

Ex. 2.10.8

Show that the dual space of the space c_0 is l^1.

Solution. We need to show that $\exists T : c_0' \longmapsto l^1$, that is (i) linear, (ii) continuous, (iii) norm preserving, (iv) 1–1, and (v) "onto".

Defiine $Tf = (f(e_1), f(e_2), \ldots)$ where (e_k) is a Schauder basis for c_0 and $f \in (c_0)'$.

(i) linear Let $\alpha, \beta \in K$, and $f, g \in c_0$.

$$\begin{aligned} T(\alpha f + \beta g) &= \big((\alpha f + \beta g)(e_1), (\alpha f + \beta g)(e_2), \dots\big) = \\ &= \big(\alpha f(e_1), \alpha f(e_2), \dots\big) + \big(\beta g(e_1), \beta g(e_2), \dots\big) = \\ &= \alpha\big(f(e_1), f(e_2), \dots\big) + \beta\big(g(e_1), g(e_2), \dots\big) = \\ &= \alpha Tf + \beta Tg, \quad \square. \end{aligned} \tag{2.0.6}$$

(ii) continuous Let $x = (\alpha_1 e_1, \alpha_2 e_2, \ldots) \in c_0$ and note that $\lim_{n\to\infty} \alpha_n = 0$. For $s_n = \sum_{k=1}^{n} \alpha_k e_k$, we have $s_n \to x$.

Let $f \in (c_0)'$. Then by the continuity of f, we have $f(s_n) \to f(x)$.

Since x is arbitray, this gives us that Tf is continuous on c_0. Similarly, since f is arbitray, this gives us that T is continuous on c_0'.

(iii) norm preserving Let $x = (\alpha_1 e_1, \alpha_2 e_2, \ldots) \in c_0$ and $f \in (c_0)'$.

Then

$$\begin{aligned} |f(x)| = |\sum_{i=1}^{\infty} \alpha_i f(e_i) &\leq \sum_{i=1}^{\infty} |\alpha_i f(e_i)| \leq \\ &\leq \sup_{n\in\mathbb{N}} |\alpha_n \left| \sum_{i=1}^{\infty} |f(e_i)| \right| = ||x||_\infty \circ \sum_{i=1}^{\infty} |f(e_i)|, \end{aligned} \tag{2.0.7}$$

hence

$$||f|| = \sup_{\substack{x\in c_0 \\ ||x||=1}} |f(x)| \leq 1 \circ \sum_{i=1}^{\infty} |f(e_1)| = ||Tf||_1. \tag{2.0.8}$$

Now consider

$$x^{(n)} = \left(\frac{|f(e_1)|}{f(e_1)}, \frac{|f(e_2)|}{f(e_2)}, \ldots, \frac{|f(e_n)|}{f(e_n)}, 0, 0, \ldots \right), \tag{2.0.9}$$

unless $f(e_k) = 0$, in which case the kth component is set to zero.

We know $x^{(n)} \in c_0$ for each $n \in \mathbb{N}$ as each has finite many nonzero terms. Note that $||x^{(n)}||_\infty = 1$ for each n. Then

$$\begin{aligned} |f(x^{(n)})| &= \left| \frac{|f(e_1)|}{f(e_1)} f(e_1) + \frac{|f(e_2)|}{f(e_2)} f(e_2) + \ldots + \frac{|f(e_n)|}{f(e_n)} f(e_n) \right| \\ &= |f(e_1)| + |f(e_2)| + \ldots + |f(e_n)|, \end{aligned} \tag{2.0.10}$$

and so

$$\begin{aligned} ||Tf||_1 = \sum_{i=1}^{\infty} |f(e_i)| &= \lim_{n\to\infty} |f(x^{(n)})| \\ &\geq \lim_{n\to\infty} \left(||f|| \, ||x^{(n)}||_\infty \right) = ||f||. \end{aligned} \tag{2.0.11}$$

Use (2.0.8) together with (2.0.11), and that gives us that $||Tf|| = ||f||$.

(iv) 1–1 Let $f \in \mathcal{N}(T)$. Then $Tf = 0$ and $||Tf|| = ||f|| = 0$.

$f = 0$, hence T is 1–1.

(v) "onto" Let $(\beta_1, \beta_2, \ldots) \in l^1$ and define $g : c_0 \longmapsto \mathbb{R}$ by $g(e_n) = \beta_n$, for each $n \in \mathbb{N}$.

Extend this linearity to $g(x) = \sum_{k=1}^{\infty} \alpha_k g(e_k) = \sum_{k=1}^{\infty} \alpha_k \beta_k$, where, $x = (\alpha_1 e_1, \alpha_2 e_2, \ldots) \in c_0$.

We need to verify that $g \in (c_0)'$: clearly g is linear, and to show that g is also bounded we have

$$|g(x)| = |\sum_{k=1}^{\infty} \alpha_k \beta_k| \leq \sup_{n \in \mathbb{N}} |\alpha_n| \sum_{k=1}^{\infty} |\beta_k| < +\infty. \tag{2.0.12}$$

We have $g \in (c_0)'$ and $Tg = (\beta_1, \beta_2, \ldots)$, hence T is onto.

All (i-v) together, we can conclude that the dual space of the space c_0 is l^1.

Chapter 3

Inner product spaces

Ex. 3.1.7

If in an inner product space, $\langle x, u\rangle = \langle x, v\rangle$ for all x, show that $u = v$.

Solution

Let $x = u - v$, then

$$\langle x, u\rangle - \langle x, v\rangle = \langle x, u - v\rangle = \langle u - v, u - v\rangle = 0.$$

This implies $u - v = 0$ and so $u = v$.

Ex. 3.1.11

Let $\mathbf{X}$ be the vector space of all ordered pairs of complex numbers. Can we obtain the norm defined on $\mathbf{X}$ by

$$\|x\| = \mid \xi_1 \mid + \mid \xi_2 \mid \qquad [x = (\xi_1, \xi_2)]$$

from an inner product?

Solution

We check whether or not the parallelogram equality is satisfied. Let $x = x(1,0)$ and $y = (0,1)$.

$$\begin{aligned} \|x + y\|^2 &= \|(1,1)\|^2 = 4. \\ \|x - y\|^2 &= \|(1,-1)\|^2 = 4. \\ \|x\|^2 = 1, \|y\|^2 &= 1. \\ \|x + y\|^2 + \|x - y\|^2 &= 8 \neq 2(\|x\|^2 + \|y\|^2) = 4. \end{aligned} \tag{3.0.1}$$

Therefore we cannot obtain the norm defined on $\mathbf{X}$ from an inner product.

Ex. 3.2.5

Show that for a sequence (x_n) in an inner product space the conditions $\|x_n\| \to \|x\|$ and $\langle x_n, x\rangle \to \langle x, x\rangle$ imply convergence $x_n \to x$

Solution

$$
\begin{aligned}
\|x_n - x\|^2 &= \langle x_n - x, x_n - x\rangle, \\
&= \langle x_n, x_n\rangle - \langle x_n, x\rangle - \langle x, x_n\rangle + \langle x, x\rangle, \\
&= \|x_n\|^2 + \|x\|^2 - \langle x_n, x\rangle - \langle x, x_n\rangle, \\
&= \|x\|^2 + \|x\|^2 - \langle x, x\rangle - \langle x, x\rangle, \text{ as } n \to \infty, \\
&= 2\|x\|^2 - 2\|x\|^2, \text{ as } n \to \infty, \\
&= 0, \text{ as } n \to \infty.
\end{aligned}
\tag{3.0.2}
$$

Therefore $x_n \to x$.

Ex. 3.3.2

Show that the subset $M = \{y = (\eta_j) | \sum_j \eta_j = 1\}$ (i) of complex space $\mathbb{C}$ is complete and convex. (ii) Find the vector of minimum norm in M.

Solution

Completness We known $||x||_1 = |\xi_1| + \ldots + |\xi_n|$ where $x = (\xi_j)$ is a norm in $\mathbb{C}^n$.

Since every norm on a finite dimensional vector space produces the same topology, it is enough to show that M is complete with respect to $||\cdot||_1$.

Let (x_m) be a Cauchy sequence in M. Since $\mathbb{C}^n$ is compelte, there exists a vector $x = (\xi_1, \xi_2, \ldots, \xi_n)$ s.t., $x_n \to x$.

We know that $\sum_j \xi_j^m = 1$ for all m. So,

$$
\begin{aligned}
|1 - \sum_j \xi_j| &\leq |1 - \sum_j \xi_j^m| + |\sum_j \xi_j^m - \sum_j \xi_j| \\
&\leq |\sum_j \xi_j^m - \sum_j \xi_j| = ||x_m - x||_1 \to 0.
\end{aligned}
\tag{3.0.3}
$$

M is closed, so it is complete in $\mathbb{C}^n$.

Convexity Let $x = (\xi_j)$, $y = (\eta_j)$, $x, y \in M$, and $\alpha \in [0, 1]$.

Since

$$
\sum_j \Big(\alpha\xi_j + (1-\alpha)\eta_j\Big) = \alpha \sum_j \xi_j + (1-\alpha)\sum_j \eta_j = \alpha + (1-\alpha) = 1, \tag{3.0.4}
$$

we obtain $\alpha x + (1-\alpha)y \in M$, so M is convex.

(ii) Let $x = (\xi_j) \in M$. Then by Hölder inequality

$$\frac{1}{\sqrt{n}} = \frac{1}{\sqrt{n}} \sum_j^n \xi_j \le \frac{1}{\sqrt{n}} \sum_j^n |\xi_j| = \sum_j^n \frac{|\xi_j|}{\sqrt{n}}$$
$$\le \left(\sum_j^n \frac{1}{\sqrt{n}^2} \right)^{1/2} \left(\sum_j^n \xi_j^2 \right)^{1/2} \tag{3.0.5}$$
$$= \left(\sum_j^n \xi_j^2 \right)^{1/2} = ||x||_2,$$

i.e., vector of M cannot have a norm less than $\frac{1}{\sqrt{n}}$. But, $y = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) \in M$, has norm $||y||_2 = \frac{1}{\sqrt{n}}$, so the vector of M with minimum norm, is actually $y = (\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n})$.

Ex. 3.3.4

(a) Show that the conclusion of Theorem 3.3-1 (Minimizing vector) also holds if X is a Hilbert space and $M \subset X$ is a closed subspace. (b) How could we use Appolonious' identity in the proof of Theorem 3.3-1?

Answer:
(a) Since X is a Hilbert space (and thus a Banach space), by Theorem 2.3-1 (Subspace of a Banach space) M is complete if and only if M is closed. Therefore the conclusion of Theorem 3.3-1 follows immediately. (b) Appolonious' identity can be used when proving existence and uniqueness, replacing the use of the parallelogram equality. To prove existence we show that a given sequence $(y_n) \in M$ such that $\delta_n \longrightarrow \delta$ where $\delta_n = \|x - y_n\|$, is Cauchy. Here δ denotes the distance of $x \in X$ to the minimizing vector $y \in M$, i.e. $\|x - y\| = \delta$. By Appolonious' identity we have,

$$\|x - y_n\|^2 + \|x - y_m\|^2 = \frac{1}{2}\|y_n - y_m\|^2 + 2\|x - \frac{1}{2}(y_n + y_m)\|^2 \Longleftrightarrow$$
$$\|y_n - y_m\|^2 \le 2\delta_n^2 + 2\delta_m^2 - 4\delta^2.$$

In the inequality we used the convexness of M to obtain the bound
$\|x - \frac{1}{2}(y_n + y_m)\| \le \delta$. Since $\delta_n, \delta_m \longrightarrow \delta$, for all $\varepsilon > 0$ there exists $N > 0$ such that for all $m, n > N$ we have $\|y_n - y_m\| \le 2\delta_n^2 + 2\delta_m^2 - 4\delta^2 < \varepsilon$ which shows that (y_n) is Cauchy. Since M is complete we have $y_n \longrightarrow y \in M$, which shows existence.

To prove uniqueness we show that given $y, y_0 \in M$, satisfying
$\|x - y\| = \|x - y_0\| = \delta$, we must have $y = y_0$. Appolonious' identity then gives

$$\|x - y\|^2 + \|x - y_0\|^2 = \frac{1}{2}\|y - y_0\|^2 + 2\|x - \frac{1}{2}(y + y_0)\|^2 \Longleftrightarrow$$
$$\|y - y_0\|^2 \le 2\delta^2 + 2\delta^2 - 4\delta^2 = 0,$$

where again the convexness of M was used to bound $\|x - \frac{1}{2}(y + y_0)\| \leq \delta$. Since the norm is non-negative $y = y_0$ follows, proving uniqueness.

Ex. 3.4.6

Let $\{e_1, \ldots, e_n\}$ be an orthonormal set in an inner prouct space X, where n is fixed. Let $x \in X$ be any fixed element and $y = \beta_1 e_1 + \ldots + \beta_n e_n$. Then $||x - y||$ dpeends on $\beta_1, \ldots, \beta_n$. Show by direct calculation that $||x - y||$ is minimum iff $\beta_j = \langle x_i, e_i \rangle$, where $j = 1, \ldots, n$.

Solution Consider,

$$
\begin{aligned}
||x-y||^2 = \langle x-y, x-y \rangle = \langle x - \sum_j^n \beta_j e_j, x - \sum_j^n \beta_j e_j \rangle = \\
&= \langle x, x \rangle - \langle x, \sum_j^n \beta_j e_j \rangle - \langle \sum_j^n \beta_j e_j, x \rangle + \langle \sum_j^n \beta_j e_j, \sum_j^n \beta_j e_j \rangle = \\
&= ||x||^2 - \sum_j^n \bar{\beta}_j \langle x, e_j \rangle - \sum_j^n \beta_j \langle e_j, x \rangle + \sum_j^n \bar{\beta}_j \beta_j = \\
&= ||x||^2 + \sum_j^n \Big(- \bar{\beta}_j \langle x, e_j \rangle - \beta_j \langle e_j, x \rangle + \bar{\beta}_j \beta_j \Big) = \ldots
\end{aligned}
\tag{3.0.6}
$$

Add and substract $\sum_j^n \langle x, e_j \rangle \overline{\langle x, e_j \rangle}$,

$$
\begin{aligned}
\ldots &= ||x||^2 + \sum_j^n \Big(- \bar{\beta}_j \langle x, e_j \rangle - \beta_j \langle e_j, x \rangle + \bar{\beta}_j \beta_j + \langle x, e_j \rangle \overline{\langle x, e_j \rangle}, \Big) - \sum_j^n \langle x, e_j \rangle \overline{\langle x, e_j \rangle} = \\
&= ||x||^2 + \sum_j^n (\bar{\beta}_j - \langle x, e_j \rangle)(\beta_j - \langle e_j, x \rangle) - \sum_j^n |\langle x, e_j \rangle|^2 = \\
&= ||x||^2 + \sum_j^n |\beta_j - \langle x, e_j \rangle|^2 - \sum_j^n |\langle x, e_j \rangle|^2.
\end{aligned}
\tag{3.0.7}
$$

Now, since $x, e_1, \ldots, e_n$ are fixed. Clearly, the minimum is found when $\beta_j = \langle x, e_j \rangle$, where $j = 1, \ldots, n$.

Ex. 3.4.8

Show that an element x of an inner product space $\mathbf{X}$ cannot have "too many" Fourier coefficients $\langle x, e_k \rangle$ which are "big"; here, (e_k) is a given orthonormal sequence; more precisely, show that the number n_m of $\langle x, e_k \rangle$ such that $| \langle x, e_k \rangle | > \frac{1}{m}$ must satisfy $n_m < m^2 \|x\|^2$

Solution

From Bessel's inequality, we have

$$\sum_{k=1}^{\infty} | \langle x, e_k \rangle |^2 \leq \|x\|^2.$$

Summing over those elements such that $| \langle x, e_k \rangle | > \epsilon$, then

$$N(\epsilon)\epsilon^2 \leq \sum_{k:|\langle x, e_k \rangle|>\epsilon}^{\infty} 1 \leq \sum_{k=1}^{\infty} | \langle x, e_k \rangle |^2 \leq \|x\|^2. \tag{3.0.8}$$

Therefore

$$N(\epsilon) < \frac{\|x\|^2}{\epsilon^2}.$$

Ex. 3.5.4

We shall show that if (x_j) is a sequence in an inner product space X such that

$$\sum_{j=1}^{\infty} \|x_j\| \quad \text{converges,} \tag{3.0.9}$$

then (s_n) given by $s_n = x_1 + \ldots + x_n$ is Cauchy.

Assume without loss of generality that $n > m$, and note that

$$\|s_n - s_m\| = \left\| \sum_{j=m+1}^{n} x_j \right\| \leq \sum_{j=m+1}^{n} \|x_j\|, \tag{3.0.10}$$

by the triangle inequality. Since (3.0.9) converges we know that for every $\varepsilon > 0$ there is an N such that

$$\sum_{j=k+1}^{\infty} \|x_j\| < \varepsilon \qquad \text{if } k > N. \tag{3.0.11}$$

From (3.0.10) it now follows that

$$\|s_n - s_m\| \leq \sum_{j=m+1}^{n} \|x_j\| \leq \sum_{j=m+1}^{\infty} \|x_j\| < \varepsilon \qquad \text{if } n > m > N. \tag{3.0.12}$$

This shows that (s_n) is Cauchy.

Ex. 3.5.4

Show that if $\{x_n\}$ is a sequence in an inner product space X s.t. $\sum\limits_{i=1}^{+\infty} \|x_i\| < +\infty$, then $\{s_n\}$ is a Cauchy sequence, where $s_n = \sum\limits_{i=1}^{n} x_i$.

Solution. We need to demonstrate that

$$\forall \varepsilon > 0\ \exists N \in \mathbb{N}:\ \forall n, m > N \ \Rightarrow \|s_m - s_n\| < \varepsilon.$$

By definition, $s_n = \sum\limits_{i=1}^{n} x_i,\ s_m = \sum\limits_{i=1}^{m} x_i$, thus $s_m - s_n = \sum\limits_{i=n+1}^{m} x_i$. Then

$$\|s_m - s_n\| = \left\| \sum_{i=n+1}^{m} x_i \right\| \leq \sum_{i=n+1}^{m} \|x_i\| = S_m - S_n,$$

where S_m is a partial sum of $\sum\limits_{i=1}^{n} \|x_i\|$. This series converges, thus the sequence of its partial sums $\{S_n\}$ has a limit, thus $\{S_n\}$ is Cauchy and thus

$$\forall \varepsilon > 0\ \exists N \in \mathbb{N}:\ \forall n, m > N \ \Rightarrow \|s_m - s_n\| \leq \|S_m - S_n\| < \varepsilon.$$

□

Ex. 3.5.6

Let (e_j) be an orthonormal sequence in a Hilbert space H. Show that if

$$x = \sum_{j=1}^{\infty} \alpha_j e_j, \quad \sum_{j=1}^{\infty} \beta_j e_j, \text{ then } \langle x, y \rangle = \sum_{j=1}^{\infty} \alpha_j \bar{\beta}_j$$

the series being absolutely convergent.

Solution Let $x_n = \sum_{j=1}^{n} \alpha_j e_j$ and $y_n = \sum_{j=1}^{n} \beta_j e_j$. Then $x_n \to x$ and $y_n \to y$.
Since (e_j) is an orthonormal sequence we get

$$\langle x_n, y_n \rangle = \langle \sum_{j=1}^{n} \alpha_j e_j, \sum_{j=1}^{n} \beta_j e_j \rangle = \sum_{j=1}^{n} \alpha_j \bar{\beta}_j. \qquad (3.0.13)$$

But since we know that the inner product is continous, that is $x_n \to x$ and $y_n \to y$, it implies $\langle x_n, y_n \rangle \to \langle x, y \rangle$. Hence

$$\langle x_n, y_n \rangle \to \sum_{j=1}^{\infty} \alpha_j \bar{\beta}_j = \langle x, y \rangle. \qquad (3.0.14)$$

Absolute convegence follows from, using Cauchy-Schwarz and Bessel inequalities

$$\begin{aligned}\sum_{j=1}^{\infty} |\langle x, e_j\rangle\langle y, e_j\rangle| \leq \left(\sum_{j=1}^{\infty} |\langle x, e_j\rangle|^2\right)^{\frac{1}{2}} \left(\sum_{j=1}^{\infty} |\langle y, e_j\rangle|^2\right)^{\frac{1}{2}} \\ \leq ||x||\,||y||\end{aligned} \tag{3.0.15}$$

$\square$.

Ex. 3.6.10

Show that if $M \subset H$, H is a Hilbert space and $\forall x \in M$, $\forall v, w \in H$ $\langle v, x\rangle = \langle w, x\rangle \Rightarrow v = w$, then M is total in H.

Solution. One can exploit theorem 3.6-2 to show this. It says that M is total in H if

$$v \perp M \Rightarrow v = 0.$$

Consider $v \perp M$. It means that $\forall x \in M \langle v, x\rangle = 0$. At the same time, since $\langle v, x\rangle = \langle w, x\rangle \Rightarrow v = w$ $\forall v, w \in H$ we have

$$0 = \langle v, x\rangle = \langle 0, x\rangle \Rightarrow v = 0.$$

$\square$

Ex. 3.9.2

Show that $(T^*)^{-1}$ exists and

$$(T*)^{-1} = (T^{-1})^*,$$

where H is a Hilbert space, $T : H \to H$ is a bijective bounded linear operator whose inverse is bounded.

Solution. By theorem 3.9-2 operator T^* exists, and it is unique and bounded. Operators T^{-1} and $(T^*)^{-1}$ also exist and are unique by bounded inverse theorem. Then, $\forall x, y \in H$ we have

$$\langle x, y\rangle = \langle Ix, y\rangle = \langle TT^{-1}x, y\rangle = \langle T^{-1}x, T^*y\rangle = \langle x, (T^{-1})^*T^*y)\rangle,$$

thus $I = (T^{-1})^*T^*$ and $(T^*)^{-1} = (T^{-1})^*$. $\square$
