Student’s Book
Numerical Functional Analysis

Editor: Stefan Engblom

2022
Preface

This portfolio collects the student’s output during the course *Numerical Functional Analysis*, which was given for the third time in the spring 2022 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 as usual very 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 Computational problems, 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, June 2022
Contents

I  Student’s essays 3

Fixed point proof for existence and uniqueness of Stochastic Differential Equations with Lipschitz coefficients
*Erik Blom* 4

Thermodynamic consistency of the MHD equations in presence of magnetic monopoles
*Tuan Anh Dao* 11

The adjoint-state method for gradient computations
*Gustav Eriksson* 18

Measures of Statistical Difference in Bayesian Inference
*Jonas Evaeus* 24

Solvability of integral equations on Lipschitz domains
*David Krantz* 29

Weak Convergence and Its Properties
*Gesina Menz* 37

Nonlinear Control meets Functional Analysis
*Anh Tung Nguyen* 43

Stability analysis with the Koopman and Perron-Frobenius operators
*Håkan Runvik* 50

Sobolev inner-product spaces with applications to finite element methods
*Ivy Weber* 56

II  Solutions to exercises 65

1  Metric spaces 66
2 Normed spaces 77
3 Inner product spaces 86
Part I

Student’s essays
Fixed point proof for existence and uniqueness of Stochastic Differential Equations with Lipschitz coefficients

Erik Blom*

June 17, 2022

Abstract

We use the Banach fixed point theorem to show existence and uniqueness of first order stochastic differential equations whose coefficients are Lipschitz continuous and bounded. A brief overview of the Banach fixed point theorem and the theory of stochastic processes and equations is presented followed by the proof, which combines elements of the proof for Picard’s theorem for ordinary differential equations presented in Kreyszig [1991] and essential ideas in a proof for stochastic differential equations found in Barth and Kussmaul [1981].

1 Banach Fixed Point Theorem

We begin by stating the Banach fixed point theorem, followed by a review of the relevant concepts in the field of stochastic processes and differential equations used in the final proof.

Using notation similar to what is used in Kreyszig [1991], we state the definition of a contraction mapping.

**Definition 1.1** (Contraction). Let $X = (X,d)$ be a metric space. Then a mapping $T : X \to X$ on $X$ for which there exists an $\alpha < 1$ such that

$$d(Tx, Ty) \leq \alpha d(x, y), \quad \forall x, y \in X,$$

(1.1)

is called a contraction on $X$.

*Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden. erik.blom@it.uu.se*
Furthermore, a fixed point of a mapping $T : X \rightarrow X$ is an $x \in X$ for which $Tx = x$. Fixed points and contraction mappings are essential in the Banach Fixed point theorem which can be found in Kreyszig [1991], and which states the following.

**Theorem 1.1** (Banach Fixed Point Theorem). Let $X = (X, d)$ be a complete, non-empty metric space. Then a contraction mapping $T : X \rightarrow X$ on $X$ has precisely one fixed point.

This important theorem can be used for, e.g., showing existence and uniqueness of first order ordinary differential equations as is shown in Kreyszig [1991], and we will in our proof use similar ideas for the existence and uniqueness of stochastic differential equations.

## 2 Stochastic Processes

Before proceeding to the proof we wish to informally establish the context of probability spaces and stochastic processes using the notation from Øksendal [2003]. Such processes may be solutions to stochastic differential equations (SDEs) such as eq. (3.1) presented below, and have a wide range applications in topics including noise filtering, diffusion theory, stochastic control, mathematical finance, and modelling chemical and biological systems.

For a general random process, let $\Omega$ be the sample space of all possible outcomes, $\mathcal{F}$ the space of events, or set of subsets of $\Omega$ (the $\sigma$-algebra of $\Omega$). The probability measure $P$ on the measurable space $(\Omega, \mathcal{F})$ maps each event to its corresponding probability. The resulting probability space $(\Omega, \mathcal{F}, P)$ is in this article assumed to be complete. We consider one dimensional random variables, $X$, in a probability space with $L^p(\Omega) = \{X(\omega) \in \mathbb{R}, \|X\|_p < \infty\}$, where

$$\|X\|_p = \left( \int_{\Omega} |X(\omega)|^p dP(\omega) \right)^{\frac{1}{p}},$$

for $p \in [1, \infty)$ (for the purposes of this proof, a norm defined for finite $p$ will suffice) and $\omega \in \Omega$ is a convenient label for any given 'experiment' on the stochastic process whose outcome is $X(\omega)$. The expectation value then satisfies $E[|X|^2] = \|X\|_2^2$ for a random variable $X \in L^2(\Omega)$. The expectation value is a norm up to stochastic equivalence, i.e., two stochastic processes, $X$ and $Y$ are equivalent, i.e., $P(X(\omega) = Y(\omega); \omega) = 1$, and thus these spaces are complete normed spaces, Banach spaces.

The proof in this article uses some essential properties of the stochastic integral, or Itô integral.
\[
\int_0^t f(t, \omega) dB_s(\omega), \tag{2.2}
\]

where \(dB_s(\omega)\) is 1-dimensional Brownian motion. For the Itô integral it is possible to show the following Øksendal [2003]

**Theorem 2.1.** Let \(f \in \mathcal{V}(0,T)\). Then there exists a \(t\)-continuous stochastic process, \(J_t\), on the probability space \((\Omega, \mathcal{F}, P)\), such that

\[
J_t(\omega) = \int_0^t f(t, \omega) dB_s(\omega), \quad \text{almost all } \omega. \tag{2.3}
\]

Here the functions in \(\mathcal{V}(0,T)\) satisfy certain properties such as adaption to a filtration with regards to the Brownian motion, and square integrability with respect to time of their expected value. Furthermore, following the definition in Øksendal [2003], we note that a stochastic process is a parametrized collection of random variables - the set \(\{X_t\}_{0 \leq t \leq T}\) - defined on \((\Omega, \mathcal{F}, P)\) which assumes real values. A normed space of stochastic processes thus motivate the use of a different norm than that the space of random variables, as we will see in the proof in the next section. Finally, we simply state the useful isometry of Itô integrals between different spaces, known as the Itô isometry:

\[
E[\left(\int_0^T \sigma(t, X_s) dB_s\right)^2] = E[\int_0^T |\sigma(t, X_s)|^2 ds], \tag{2.4}
\]

which presents a convenient way of converting the integration variable from the troublesome Brownian noise differential to integration with respect to time. We are now ready to understand the main ideas behind the theorem in the following section and its proof.

### 3 Existence and uniqueness proof

What this article intend to show is the following, combining ideas from the elegant proofs for ordinary differential equations in Kreyszig [1991] and SDEs Barth and Kussmaul [1981].

**Theorem 3.1.** Consider the stochastic differential equation

\[
dX_t = b(t, X_t)dt + \sigma(t, X_t)dB_t, \quad 0 \leq t \leq T, \quad X_0 = Z, \tag{3.1}
\]

where \(Z\) is independent of the \(\sigma\)-algebra generated by the Brownian process \(B_t\) with
and the coefficients of the differential equations are joint Lipschitz

\[ |b(t, x) - b(t, y)| + |\sigma(t, x) - \sigma(t, y)| \leq D|x - y| \quad (3.3) \]

and for some \( C > 0 \), bounded by

\[ |b(t, x)| + |\sigma(t, x)| \leq C(1 + |x(t)|). \quad (3.4) \]

Then eq. (3.1) has a unique \( t \)-continuous solution \( X_t(\omega) \), given that it is filtrated by \( Z \) and \( B_s \) for \( s \leq t \).

The main idea of the following proof is to introduce a mapping whose fixed points are solutions to the SDE, and show that this mapping maps to and from complete spaces. By showing that such a mapping is a contraction, the Banach fixed point theorem completes the proof.

Filtration of the solution \( X_t(\omega) \) with regards to the Brownian motion ensures that there exists a \( t \)-continuous version of the integral form of (3.1):

\[
X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dB_s, \quad \text{a.a} \ \omega. \quad (3.5)
\]

by Theorem 2.1 applied to the Itô integral on the right-hand side. To apply Theorem 1.1 on the space of stochastic processes, we require a complete metric space. Thus, we let \( \tilde{C}[0,T] \) be the normed space (from which a metric can be induced) of time-continuous random variables with norm

\[
\|X_t\|_{\tilde{C}} = (\sup_{t \geq 0} (E[|X_t|^2]))^{\frac{1}{2}}, \quad (3.6)
\]

which can be shown is a norm given that \( E[|X_t|^2]^{1/2} \) is a norm, and that \( X_t \) is \( t \)-continuous (again up to stochastic equivalence). We then define a mapping \( S : \tilde{C} \rightarrow \tilde{C} \) defined by the integral form of the stochastic differential equation

\[
SX_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dB_s, \quad (3.7)
\]

where the space \( \tilde{C} \) is the subset of \( C[0,T] \) with

\[
E[|X_t - X_0|^2] \leq 4T(1 + T)C^2e^{4T(1+T)C^2} =: G, \quad (3.8)
\]

which is a condition necessary for the mapped values to not ‘stray too far’ given the bounds of the coefficients. From (3.8) we see that \( \|X_t - X_0\|_{\tilde{C}} \leq G^{\frac{1}{2}} \)
holds for all elements $X_t$ in $\hat{C}$. The induced metric can be shown to reveal that $\hat{C}$ is closed in the complete space $C[0, T]$ and therefore complete in itself (by Theorem 1.4-7 in Kreyszig [1991]). As such, Banach fixed point theorem may be applied to the mapping $S$ if we can show that it is a contraction. Furthermore, we see that $S$ properly maps onto $\hat{C}$ by

$$E[|SX_t - X_0|^2] = E[(\int_0^t b ds + \int_0^t \sigma dB_s)^2]$$

(3.9)

$$\leq 2TE[\int_0^t |b|^2 ds] + 2E[\int_0^t |\sigma|^2 ds]$$

(3.10)

$$\leq 4T(1 + T)C^2(1 + \int_0^t E[|SX_t - X_0|^2])ds$$

(3.11)

where the relation $(a + b)^2 \leq 2(a^2 + b^2)$ was used three times, once to separate to the two integrals and twice to deal with the bounds of $|b|^2$ and $|\sigma|^2$, respectively. The Itô isometry (eq. 2.4) was applied to the term with the Itô integral, and Hölder’s inequality was used for the other term whence the factor $T$ comes from $t \leq T$. Finally, Grönwall’s lemma applied to the inequality yields the condition $E[|SX_t - X_0|^2] \leq G$.

We show that the mapping $S$ is a contraction, and begin by considering two different stochastic processes $X_t$ and $Y_t$ which satisfy $X_0 = Y_0$ (i.e., they start from the same state). Let $b_X, b_Y$ and $\sigma_X, \sigma_Y$ be their respective coefficients and define $a(t) = b_X(t, X_t) - b_Y(t, Y_t)$, and $\gamma(t) = \sigma_X(t, X_t) - \sigma_Y(t, Y_t)$. Then, by a similar reason as above, we obtain that

$$E[(SX_t - SY_t)^2] = E[(X_0 - Y_0 + \int_0^t ads + \int_0^t \gamma dB_s)^2]$$

(3.12)

$$\leq 2TE[\int_0^t |a|^2 ds] + 2E[\int_0^t |\gamma|^2 ds].$$

(3.13)

We further bound this by using the square of the Lipschitz condition which bounds the difference of the coefficients by the difference between the random variables, and thus

$$2TE[\int_0^t |a|^2 ds] + 2E[\int_0^t |\gamma|^2 ds]$$

(3.14)

$$\leq 2(1 + T) \int_0^t E[(X_s - Y_s)^2 D^2]ds,$$

(3.15)

where the order of integration was also interchanged.
A key part of this proof is to find a viable metric for the use of the fixed point theorem, which motivates the following. Similar to the proof by Barth and Kussmaul [1981] we insert $e^{-\alpha D^2 s} e^{\alpha D^2 s}$ for some $\alpha \geq 0$, and obtain a useful bound

$$2(1 + T) \int_0^t E[(X_s - Y_s)^2 e^{-\alpha D^2 s} e^{\alpha D^2 s} D^2] ds$$

(3.16)

$$\leq 2(1 + T) \sup_{0 \leq s \leq t} (E[(X_s - Y_s)^2 e^{-\alpha D^2 s}]) \int_0^t e^{\alpha D^2 s} D^2 ds,$$

(3.17)

The integral is now readily solved, and after a few simple manipulations similar to the ones in the article, we arrive at

$$\sup_{0 \leq s \leq t} (E[(SX_s - SY_s)^2 e^{-\alpha D^2 s}])$$

(3.18)

$$\leq \frac{2(1 + T)}{\alpha} \sup_{0 \leq s \leq t} (E[(X_s - Y_s)^2 e^{-\alpha D^2 s}]).$$

(3.19)

Finally, we introduce the norm

$$\|X_t\|_{C^*} = (\sup_{0 \leq t \leq T} (E[|X_t|^2 e^{-\alpha D^2 t}]))^{\frac{1}{2}},$$

(3.20)

which can be shown to be equivalent to the norm of $\tilde{C}$. The metric corresponding to this norm is

$$d(x(s), y(s)) = (\sup_{0 \leq s \leq t} (E[(x(s) - y(s))^2 e^{-\alpha D^2 s}]))^{\frac{1}{2}}$$

(3.21)

Then eq. (3.19) takes the form

$$d(SX_t, SY_t)^2 \leq \frac{2(1 + T)}{\alpha} d(X_t, Y_t)^2,$$

(3.22)

and since $\alpha$ was arbitrary (and nonnegative), we may always find one $\alpha$ such that $\frac{2(1 + T)}{\alpha} < 1$, and thus the mapping $S$ is a contraction on $\tilde{C}$ (by Definition 1.1). Then, by completeness of the space, the Banach fixed point theorem states that $S$ has precisely one fixed point. That is, there is precisely one stochastic process $X_t$ that satisfies eq. (3.5) (up to stochastic equivalence of processes on the probability space, i.e., $P(\{X_t(\omega) = Y_t(\omega)\}; \omega) = 1$), and given that $X_t$ is $t$-continuous, it also satisfies the SDE (3.1). The proof for Theorem 3.1 is complete, and we may safely begin solving such interesting and useful stochastic differential equations.
References


Thermodynamic consistency of the MHD equations in presence of magnetic monopoles

Tuan Anh Dao∗

June 9, 2022

Abstract

In this essay, we investigate some thermodynamic and numerical properties of the magnetohydrodynamics (MHD) equations presuming the existence of magnetic monopoles. When magnetic monopoles are present, the second law of thermodynamics does not hold for the MHD equations in the conservative form. First, we look into the second law of thermodynamics from a functional analysis perspective. Specifically, a proof of the Clausius-Duhem inequality is presented, to which the Hahn-Banach separation theorem holds the key. Then, we show that in a slightly modified form, the MHD equations satisfy the second law of thermodynamics even in presence of the magnetic monopoles.

1 Introduction

Plasma can be modeled by the system of MHD equations. Consider a spatial domain \( \Omega \subset \mathbb{R}^d, d = 1, 2, 3 \) and a temporal domain \((t_0, t_1] \subset \mathbb{R}, 0 \leq t_0 < t_1\). Let us denote \( \rho(x, t) : (\Omega, (0, T]) \rightarrow \mathbb{R} \) the density, \( \mathbf{m} : (\Omega, (0, T]) \rightarrow \mathbb{R}^d \) the momentum, \( E(x, t) : (\Omega, (0, T]) \rightarrow \mathbb{R} \) the total energy, and \( \mathbf{B} : (\Omega, (0, T]) \rightarrow \mathbb{R}^d \) the magnetic field of the plasma. We denote the solution vector \( \mathbf{U} := (\rho, \mathbf{m}, E, \mathbf{B})^\top \). An initial solution \((\rho_0(x), \mathbf{m}_0(x), E_0(x), \mathbf{B}_0(x))\) is given at time \( t = t_0 \). The dynamics of plasma can be described by the

∗Division of Scientific Computing, Department of Information Technology, Uppsala university, SE-751 05 Uppsala, Sweden. tuananh.dao@it.uu.se
MHD equations in the conservative form,
\[\begin{align*}
&\partial_t \rho + \nabla \cdot \mathbf{m} = 0, \\
&\partial_t \mathbf{m} + \nabla \cdot (\mathbf{m} \otimes \mathbf{u} + p \mathbf{I}) - \nabla \cdot \mathbf{\beta} = 0, \\
&\partial_t \mathbf{E} + \nabla \cdot (\mathbf{u}(\mathbf{E} + p)) - \nabla \cdot (\mathbf{u} \cdot \mathbf{\beta}) = 0, \\
&\partial_t \mathbf{B} + \nabla \cdot (\mathbf{u} \otimes \mathbf{B} - \mathbf{B} \otimes \mathbf{u}) = 0,
\end{align*}\]
(1.1)

where the Maxwell stress tensor is \(\mathbf{\beta} := -\frac{1}{2}(\mathbf{B} \cdot \mathbf{B}) \mathbf{I} + \mathbf{B} \otimes \mathbf{B}\).

The matrix \(\mathbf{I}\) is the identity matrix of size \(d \times d\). The short derivative notation \(\partial_t\) denotes the corresponding partial derivative \(\partial / \partial t\). The dot 
\(\cdot\) denotes the dot product. The notation 
\(\otimes\) denotes the outer product. The notation 
\(\nabla \cdot\) denotes the divergence operator.

Multiplying each equation in (1.1) with a suitable factor and adding them together, one can obtain an additional conservation equation of a variable denoted \(S\), yet to be defined, given that \(\nabla \cdot \mathbf{B} = 0\), see e.g., [Godunov, 1972],
\[\partial_t S + \nabla \cdot (\mathbf{u} S) = 0.\]
(1.2)

The quantity \(S\) is called an entropy, and (1.2) describes the entropy conservation, which defines the second law of thermodynamics for smooth solutions. Also by thermodynamics, at discontinuities, the equal sign (1.2) becomes “greater than”, which says that entropy should always increase. We will discuss this in more details in Sections 3 and 4. Without the assumption \(\nabla \cdot \mathbf{B} = 0\), we would not have (1.2). Therefore, the so-called “divergence-free condition” \(\nabla \cdot \mathbf{B} = 0\) has a physical significance, despite being decoupled from the conservative form (1.1).

There are several reasons why investigating the case \(\nabla \cdot \mathbf{B} \neq 0\) is also relevant. The particles exhibiting \(\nabla \cdot \mathbf{B} \neq 0\) are called “magnetic monopoles”. Although there has been no evidence that such particles exist in nature and their existence is usually not accepted by physicists, magnetic monopoles are still of theoretical interest. In numerical analysis, however, we have a very practical reason to be concerned. When solving (1.1) using numerical methods, numerical errors such as truncation error, are embedded into the numerical solutions including the magnetic field solution \(\mathbf{B}\). Without any treatment of the obtained \(\mathbf{B}\), in general, the divergence-free condition is violated and \(\nabla \cdot \mathbf{B} \neq 0\). It becomes problematic because the divergence error \(|\nabla \cdot \mathbf{B}|\) does not decrease by refining the discretization. Therefore, the MHD solution may not converge to the correct solution.

Now, what if \(\nabla \cdot \mathbf{B} \neq 0\)? Assuming sufficient smoothness, applying a divergence operator to the fourth equation in (1.1) gives
\[\partial_t (\nabla \cdot \mathbf{B}) = 0.\]
(1.3)
It turns out that the divergence-free condition holds for any future time \( t > t_0 \) if the initial solution \( \mathbf{B}_0 \) is divergence-free.

The following Section 2 shows an elegant proof of the second law of thermodynamics in the form of the Clausius-Duhem inequality. In Section 3, we look at a modified version of the MHD equations (1.1). Section 4 shows that even when \( \nabla \cdot \mathbf{B} \neq 0 \), the MHD equations in the modified form indeed satisfies the entropy inequality which is another way of stating the second law of thermodynamics, see e.g., [Pelkowski, 2014].

2 The Clausius-Duhem inequality

The proof in this section is a simplification of [Feinberg and Lavine, 1984].

We start by introducing several definitions. More basic concepts such as measure, topological space, metric spaces, compactness can be found in the book [Kreyszig, 1991]. In a topological space, a Borel measure is a measure that is defined on all open sets. A Hausdorff space is a topological space in which for any two distinct points, there exists two disjoint neighborhoods each containing one point. A locally convex space is a vector space equipped with a family of seminorms. The version of the Hahn-Banach separation theorem that we use is stated below.

**Theorem 2.1** (Hahn-Banach separation theorem). Let \( X \) be a locally convex Hausdorff space and \( A, B \) be two non-empty disjoint convex subsets in \( X \). If \( A \) is closed and \( B \) is compact, then there exists a linear functional \( f : X \to \mathbb{R} \) such that

\[
f(a) \leq 0 < f(b), \quad \forall a \in A, b \in B.
\]

At a fixed time, we define an admissible set of states \( \Sigma \) of the matter being considered. For example, \( \Sigma := \{ \rho \geq \epsilon, p \geq \epsilon \} \), \( \epsilon > 0 \), where \( \rho \) denotes density, and \( p \) denotes pressure. We assume that \( \Sigma \) is a compact Hausdorff space. Consider the time interval \( [t_0, t_1] \). Let \( B \subset \Sigma \). At time \( t_j \), \( m_j(B) \) denotes the mass having the states contained by \( B \). It can be seen that \( m_j(\Sigma) \) is a constant which is equal to the total mass. Consequently, we always have \( m_1(\Sigma) - m_0(\Sigma) = 0 \). Let \( q \) be a heating measure. More precisely, \( q(B) \) presents the net heat gained when the material stays in states contained by \( B \) from time \( t_0 \) to \( t_1 \). Denote \( M(\Sigma) \) the set of Borel measures on \( \Sigma \), and \( M^0(\Sigma) := \{ v \in M(\Sigma) \mid v(\Sigma) = 0 \} \). The pairs \( (m_1 - m_0, q) \) are then members of \( M^0(\Sigma) \bigoplus M(\Sigma) \). It follows that \( M(\Sigma) \) and \( M^0(\Sigma) \bigoplus M(\Sigma) \) are locally convex Hausdorff spaces. Let us say that we are using a thermodynamic theory where \( P \) is the set of all admissible pairs \( \{(\Delta m, \nu)\} \), \( \Delta m = m_1 - m_0 \). For this essay, we assume that \( P \) is a closed convex cone. We also denote
$K := \{(0, \nu) \mid \nu \geq 0, \nu(\Sigma) = 1\}$. $K$ is a compact convex set because $\Sigma$ is compact. Both $K$ and $P$ are subsets of $M^0(\Sigma) \oplus M(\Sigma)$.

**Definition 2.1** (Cyclic process). A process from $t_0$ to $t_1$ is called a cyclic process if $\Delta m = m_1 - m_0 = 0$, meaning that at time $t_1$, the amount of mass in each state is the same as it was at the initial time $t_0$.

**Definition 2.2** (Kelvin-Planck condition). In words, the Kelvin-Planck theory says that it is impossible for a cyclic process to only absorb heat from external sources but not emit heat itself. Mathematically, the convex cone $P$ is said to satisfy the Kelvin-Planck condition if $P$ contains no pairs $(0, \nu)$ where $\nu(B) \geq 0 \ \forall B \subset \Sigma, \nu(\Sigma) > 0$. Consequently, $P \cap K = \emptyset$.

**Lemma 2.2** (Riesz-Markov representation theorem). Recall that the space $M^0(\Sigma) \oplus M(\Sigma)$ is locally convex Hausdorff. For any continuous linear functional $f : M^0(\Sigma) \oplus M(\Sigma) \to \mathbb{R}$, there exist $\alpha, \beta \in C^0(\Sigma, \mathbb{R})$ such that
\[
f(v, w) = \int_\Sigma \alpha \, dv + \int_\Sigma \beta \, dw,
\]
where $v, w$ are Borel measures on $\Sigma$. The space $C^0(\Sigma, \mathbb{R})$ contains all continuous mappings from $\Sigma$ to $\mathbb{R}$.

Finally, the main result of this section is established below.

**Theorem 2.3** (Clausius-Duhem inequality). Under the Kelvin-Planck condition, there exists $S \in C^0(\Sigma, \mathbb{R})$ and $T \in C^0(\Sigma, \mathbb{R}_+)$ such that
\[
\int_\Sigma s \, dm_1 - \int_\Sigma s \, dm_0 =: \int_\Sigma s \, d(\Delta m) \geq \int_\Sigma \frac{d(q)}{T}.
\]
(2.1)
The inequality (2.1) is called the Clausius-Duhem inequality.

*Proof.* Recall that $P$ is a closed convex cone and $K$ is a compact convex set. In addition, by Definition 2.2, $P$ and $K$ are disjoint. Applying the Hahn-Banach separation theorem 2.1, there exists a linear functional $f : M(\Sigma) \oplus M(\Sigma) \to \mathbb{R}$, such that $f(P) \leq 0$, i.e.,
\[
f(\Delta m, q) \leq 0, \ \forall (\Delta m, q) \in P,
\]
(2.2)
and $f(K) > 0$, i.e.,
\[
f(0, q) > 0, \ \forall (0, q) \in K.
\]
(2.3)
From Lemma 2.2, choose $\alpha = -s$ and $\beta = \frac{1}{T}$, then $f$ can be expressed as
\[
f(\mu, \nu) = \int_\Sigma \frac{1}{T} \, d\nu - \int_\Sigma s \, d\mu
\]
We can show that $T > 0$. For each state $\sigma \in \Sigma$, the Dirac function $\delta_\sigma$ being a member of $K$ because $\delta_\sigma(\Sigma) = 1$, makes $f(0, \delta_\sigma) > 0$. This means $f(0, \nu) = \int_{\Sigma} \frac{1}{T} \mathrm{d}\nu > 0$ at any given state $\sigma$. Therefore, $T$ only takes positive values.

Set $\mu = \Delta m$ and $\nu = q$, we obtain (2.1). A pair $(T, s)$ satisfying (2.1) is said to be a Clausius-Duhem pair. The mapping $T$ is called a Clausius-Duhem temperature scale, and $s$ is then the corresponding specific entropy function.

The integrals in (2.1) are in the state space $\Sigma$. The spatial interpretation of the Clausius-Duhem inequality is

$$\frac{\mathrm{d}}{\mathrm{d}t} \int_{\Omega} \rho s \mathrm{d}V \geq \int_{\Omega} \frac{\rho e}{T} \mathrm{d}V.$$  

### 3 Godunov form of the MHD equations

Godunov [1972] introduced the following modified set of equations,

\begin{align*}
\partial_t \rho + \nabla \cdot m &= 0, \\
\partial_t m + \nabla \cdot (m \otimes m + \rho \mathbb{I}) - \nabla \cdot \beta &= -B(\nabla \cdot B), \\
\partial_t E + \nabla \cdot (u(E + p)) - \nabla \cdot (u \cdot \beta) &= -(u \cdot B)(\nabla \cdot B), \\
\partial_t B + \nabla \cdot (u \otimes B - B \otimes u) &= -u(\nabla \cdot B). \tag{3.1}
\end{align*}

The only difference to (1.1) is the right-hand-side source terms proportional to $\nabla \cdot B$. When $\nabla \cdot B = 0$, the two systems are identical. If one considers $\nabla \cdot B \neq 0$, (3.1) has several attractive advantages over (1.1), see [Powell et al., 1999]. However, the advantages come with a considerable price that the system (3.1) is no longer conservative. If the divergence error $|\nabla \cdot B|$ is large, loss of conservativeness leads to unphysical behaviors, especially when the solution contains discontinuities. Instead of (1.3) as in the $\nabla \cdot B = 0$ case, the following evolution equation of divergence is followed,

$$\frac{\partial}{\partial t} (\nabla \cdot B) + \nabla \cdot (u(\nabla \cdot B)) = 0. \tag{3.2}$$

The equation (3.2) describes the advection of non-zero divergence from the monopoles. Numerically, we expect that this advection is significant enough to prevent accumulation of divergence error at the monopoles. [Powell et al., 1999] reported that with this formulation, the divergence error $|\nabla \cdot B|$ that can be achieved is within truncation error. Rewrite the system (3.1) as

$$U_t + \nabla \cdot f(U) = r, \tag{3.3}$$
where \( f(U) := f_{\text{Euler}}(U) + f_{\text{MHD}}(U) \),

\[
\begin{align*}
    f_{\text{Euler}}(U) &:= (m, m \otimes u + pI, u(E + p), 0), \\
    f_{\text{MHD}}(U) &:= (0, -\beta, -u \cdot \beta, u \otimes B - B \otimes u), \\
    r &:= (0, -B(\nabla \cdot B), -(u \cdot B)(\nabla \cdot B), -u(\nabla \cdot B)).
\end{align*}
\]

In the next section, we will show that (3.1) indeed conserves entropy without assuming \( \nabla \cdot B = 0 \).

### 4 Conservation of entropy

The following routine is borrowed from [Bohm et al., 2020, Theorem 1].

**Theorem 4.1.** Consider an ideal gas, the modified system (3.1) exhibits an entropy function \( S \) which satisfies the following equality

\[
\int_\Omega S_t dV + \int_{\partial \Omega} (uS) \cdot n \, ds = 0,
\]

where \( n \) is the normal vector pointing outwards on \( \partial \Omega \).

**Proof.** For an ideal gas, we use the following entropy and specific entropy,

\[
S = -\frac{\rho s}{\gamma - 1}, \quad s = \ln(p\rho^{\gamma}),
\]

where \( \gamma > 0 \) is an ideal gas coefficient. Let \( \beta = \frac{\rho}{p^{\gamma}} \). We introduce the following entropy variables

\[
w = \frac{\partial S}{\partial U} = \left( \frac{\gamma - s}{\gamma - 1} - \beta u^2, 2\beta u^\top, -2\beta, 2\beta B^\top \right)^\top.
\]

Left multiplying \( w^\top \) to (3.3), we have

\[
w^\top U_t + w^\top (\nabla \cdot f_{\text{Euler}}) + w^\top (\nabla \cdot f_{\text{MHD}}) = w^\top r.
\] (4.1)

Next, we analyze each term in (4.1). The algebraic details are omitted for readability. For the first term, we have

\[
w^\top U_t = \left( \frac{\partial S}{\partial U} \right)^\top U_t = S_t.
\]

The second term has been analyzed for the compressible Euler equations,

\[
w^\top (\nabla \cdot f_{\text{Euler}}) = \nabla \cdot (uS).
\]
The third term and the right-hand-side term are equal due to the construction of the source term \( r \),
\[
\mathbf{w}^\top (\nabla \cdot \mathbf{f}^\text{MHD}) - \mathbf{w}^\top \mathbf{r} = 0.
\]
Therefore, (4.1) becomes
\[
S_t + \nabla \cdot (\mathbf{u} S) = 0.
\]
Integrating both sides over \( \Omega \), applying Gauss’s law on \( \int_{\Omega} \nabla \cdot (\mathbf{u} S) dV \), we have
\[
\int_{\Omega} S_t dV + \int_{\partial \Omega} (\mathbf{u} S) \cdot \mathbf{n} ds = 0
\]
which completes the proof.

\section{5 Conclusion}
Starting from very simple physical assumptions, we have shown why entropy should always increase. The rest of the essay motivates why a non-conservative form of the MHD equations is more suitable for numerical purposes. That is because even if there are numerical errors, the analytical entropy always increase.

\section*{References}


The adjoint-state method for gradient computations

Gustav Eriksson*

June 13, 2022

Abstract

A large number of problems in science and engineering can be formulated as equality constrained optimization problems. A flexible and efficient solution method for such problems is the adjoint-state method. In this essay, the adjoint-state method is presented for the special case of linear constraints.

1 Introduction

The problem of inferring optimal model parameters to match a desired outcome is generally formulated as a constrained minimization problem of some loss functional. Consider for example the design of aircraft wings to maximize lift, or seismic imaging to investigate subsurface conditions. A common approach for these kinds of problems is to compute the gradient of the functional (or function, depending on the model parameters), and iteratively improve the solution until a minimum is obtained. However, computing directly or approximating the gradient can be difficult. If evaluating the functional is cheap for many different model parameters, it may be feasible to approximate the gradient by essentially computing finite difference approximations in the directions of the model parameters. However, for many problems, this approach is too expensive. An alternative is to use the adjoint-state method (Plessix [2006]). This method is a general procedure for computing the gradient of a functional that depends on state variables, which in turn are solutions to a forward problem. The method introduces an adjoint-state
variable, which is given implicitly through an adjoint equation (or backward problem). By solving the forward and backward problem once, the gradient with respect to multiple model parameters can be computed relatively cheaply.

This essay presents an outline of the steps necessary to employ the adjoint-state method to a general minimization problem. In Section 2 some necessary preliminaries are introduced. In Section 3 the adjoint-state method is derived. Two examples of the method are presented in Section 4.

2 Preliminaries

To make the essay self-contained, a few necessary definitions and theorems are included in this section. Theorem 2.1 and Definition 2.1 are taken from Kreyszig [1989]. Definition 2.2 is taken from Penot [2012].

The name of the method suggests the use of adjoints, hence we shall need the following definition of a Hilbert-adjoint operator:

**Definition 2.1.** Let $T : H_1 \to H_2$ be a bounded linear operator, where $H_1$ and $H_2$ are Hilbert spaces. Then the Hilbert-adjoint operator $T^*$ of $T$ is the operator

$$T^* : H_2 \to H_1,$$

such that for all $x \in H_1$ and $y \in H_2,$

$$(Tx, y)_{H_2} = (x, T^*y)_{H_1}.$$  

(2.1)

(2.2)

For gradient-based optimization methods to make sense in the framework of vector spaces, a generalization of the standard calculus derivative is needed. Here, the Gateaux derivative is used, with the following definition:

**Definition 2.2.** Let $X$ and $Y$ be normed spaces, let $U$ be an open subset of $X$, let $f : U \to Y$ be a function, and let $x_0 \in U$. If there is some $d_{\delta_v}f(x_0) : X \to Y$ that is bounded and linear such that for all $\delta_v \in X$ we have

$$\lim_{t \to 0} \frac{f(x_0 + tv) - f(x_0)}{t} = d_{\delta_v}f(x_0),$$

then we say that $f$ is Gateaux differentiable at $x_0$ and call $d_{\delta_v}f(x_0)$ the Gateaux derivative of $f$ at $x_0$.

In Section 4.1, we consider linear and bounded functionals on Hilbert spaces. In this setting, the following theorem, often referred to as the Riesz representation theorem, is useful:
Theorem 2.1. Every bounded linear functional $f$ on a Hilbert space $H$ can be represented in terms of the inner product, namely,

$$f(x) = (z, x),$$

(2.4)

where $z \in H$ depends on $f$, is uniquely determined by $f$ and has norm

$$||z|| = ||f||.$$  

(2.5)

3 The adjoint-state method

We consider equality constrained optimization problems on the form

$$\begin{align*}
\text{minimize} & \quad J(m) = J_m(u_m), \\
\text{subject to} & \quad T_m u_m = b_m.
\end{align*}$$

(3.1)

Here $u : M \to U$, where $U$ is a Hilbert space, and $m \in M$, where $M$ is a model parameter space. The functional $J_m : U \times M \to \mathbb{R}$ defines the loss and the linear operator $T_m : U \times M \to U$ and vector $b_m : M \to U$ define the constraint. In (3.1) the $m$ dependence of $u_m$, $J_m$, $T_m$, and $b_m$ is indicated by subscripts. For the rest of this essay, the subscripts will be left out for notational clarity.

To incorporate the constraint into the minimization problem an augmented Lagrangian functional $L : U \times M \times U \to \mathbb{R}$ is constructed as

$$L(u, m, \lambda) = J(u) + (\lambda, Tu - b),$$

(3.2)

where $\lambda \in U$ is a Lagrange multiplier (or adjoint-state variable). By the theory of Lagrange multipliers, see Ciarlet et al. [1989], a solution to the constrained optimization problem (3.1) is a stationary point of the augmented Lagrangian (3.2). We proceed by computing the Gateaux derivatives of $L$ in the $u$- and $\lambda$-directions and setting them to zero, we get

$$\begin{align*}
d_{\delta u} L(u, m, \lambda) &= d_{\delta u} J(u) + (\lambda, T\delta u) = 0, \quad \forall \delta u \in U \\
d_{\delta \lambda} L(u, m, \lambda) &= (\delta \lambda, Tu - b) = 0, \quad \forall \delta \lambda \in U.
\end{align*}$$

(3.3) (3.4)

Here (3.3) is referred to as the adjoint equation and used to solve for $\lambda$. In general, it is given by

$$d_{\delta u} J(u) + (T^* \lambda, \delta u) = 0,$$

(3.5)

where Definition 2.1 has been used.
The second equation (3.4) is simply the constraint

\[ Tu = b. \quad (3.6) \]

The sought-after gradient is the derivative of \( J(m) \) with respect to the model parameter \( m \), given by

\[ d_{\delta m} J(m) = d_{\delta m} L(u, m, \lambda) = d_{\delta m} J(u) + (\lambda, d_{\delta m} (Tu - b)). \quad (3.7) \]

Computing (3.7) directly involves computing the term \( d_{\delta m} u \), which in general is difficult to evaluate since \( u \) is given implicitly by \( m \). However, by using the solution to the adjoint equation, (3.7) can be simplified so that computing this term is avoided. This is made clear by two examples in Section 4.

The procedure for computing the gradient (for example in a gradient descent iteration) is summarized in the following three steps:

1. Solve the forward problem (3.6).
2. Solve the adjoint (or backward) problem (3.5).
3. Using the obtained state and adjoint solutions \( u \) and \( \lambda \), compute the gradient according to (3.7).

4 Examples

In this section, two examples are presented to concretize the adjoint-state method.

4.1 Linear functional

Consider the special case when \( J(u) \) is linear and bounded. Then, we can use Theorem 2.1 and write

\[ J(u) = (z, u), \quad \forall u \in U, \quad (4.1) \]

where \( z \in U \). The derivative becomes

\[ d_{\delta u} J(u) = (z, \delta u), \quad \forall \delta u \in U. \quad (4.2) \]

Using (4.2), the adjoint equation (3.5) simplifies to

\[ T^* \lambda = -z. \quad (4.3) \]
If $T$ is self-adjoint, so that $T^* = T$, the adjoint equation (4.3) differs only from the constraint (3.6) by the right-hand side.

We also have

$$d_{\delta m} J(u) = (d_{\delta m} z, u) + (z, d_{\delta m} u), \quad (4.4)$$

and

$$d_{\delta m} (Tu - b) = d_{\delta m} Tu + T d_{\delta m} u - d_{\delta m} b. \quad (4.5)$$

Plugging (4.4) and (4.5) into (3.7) results in

$$d_{\delta m} J(m) = (d_{\delta m} z, u) + (\lambda, d_{\delta m} Tu - d_{\delta m} b), \quad (4.6)$$

where the adjoint equation (4.3) has been used to cancel the terms involving $d_{\delta m} u$.

### 4.2 Wave speed inversion

Consider the minimization problem

$$\min_c J(c) = J(u) = \frac{1}{2} \int_0^T r(u, t)^2 \, dt,$$

s.t.

$$u_{tt} = c^2 \Delta u + F(\bar{x}, t), \quad \bar{x} \in \Omega, \quad 0 \leq t \leq T,$$

$$u = 0, \quad \bar{x} \in \partial \Omega, \quad 0 \leq t \leq T,$$

$$u = 0, \quad \bar{x} \in \Omega, \quad t = 0,$$

$$u_t = 0, \quad \bar{x} \in \Omega, \quad t = 0, \quad (4.7)$$

where $c$ is an unknown scalar wave speed and

$$r(u, t) = u(\bar{x}_r, t) - u_r(t), \quad (4.8)$$

is the difference between the solution at $\bar{x}_r$ and known receiver data $u_r(t)$. The problem (4.7) could, for example, be used to find the wave speed in a homogeneous liquid, where the liquid is disturbed according to the forcing function $F(\bar{x}, t)$ and the pressure response $u_r(t)$ is recorded by a receiver at $\bar{x}_r$.

The domain is given by $\Omega \subset \mathbb{R}^d$ ($d = 1, 2$ or 3) with boundary $\partial \Omega$. We shall use the inner product

$$(u, v) = \int_0^T \int_\Omega u(\bar{x}, t)v(\bar{x}, t) \, d\bar{x} \, dt. \quad (4.9)$$

The constraints in (4.7) are enforced by the method of Lagrange multipliers. We have four constraints (the PDE, the boundary condition, and the
two initial conditions) and thus four adjoint-state variables are required. The details of the analysis are omitted to retain space, but the procedure follows directly from Section 3.

To obtain the adjoint equation (see (3.5)), the derivative of the misfit function with respect to the solution \( u \) is required. Here we have

\[
d_{\delta u} J(u) = \int_0^T r d_{\delta u} u(\bar{x}_r, t) = (r \delta(\bar{x} - \bar{x}_r), \delta u),
\]

where the Dirac delta function \( \delta \) is used to introduce the integral over \( \Omega \) (this allows us the write \( d_{\delta u} J(u) \) in terms of the inner product). By taking the derivative of the Lagrangian with respect to \( u \) and setting the result to zero, using integration by parts in space and time to compute the adjoint operator, and using (4.10) the adjoint equation becomes

\[
\lambda_{tt} = c^2 \Delta \lambda - r \delta(\bar{x} - \bar{x}_r), \quad \bar{x} \in \Omega, \quad 0 \leq t \leq T,
\]
\[
\lambda = 0, \quad \bar{x} \in \partial \Omega, \quad 0 \leq t \leq T,
\]
\[
\lambda_t = 0, \quad \bar{x} \in \Omega, \quad t = T.
\]

Note that the adjoint equation (4.11) is solved backward in time. If \( \lambda \) satisfies (4.11) and \( u \) the constraints in (4.7), then the derivative of \( J \) with respect to \( c \) is given by

\[
d\frac{dJ(c)}{dc} = -2c(\lambda, \Delta u).
\]

**References**


Measures of Statistical Difference in Bayesian Inference

Jonas Evaeus∗

June 19, 2022

Abstract

The landscape of the Bayesian inference problem is sketched before delving into the specific considerations of statistical difference relevant to approximate Bayesian inference. We focus on the minimization of the Kullback-Leibler (KL) divergence that is central to both variational inference and expectation propagation. After briefly discussing reasons why the KL divergence might be unsatisfactory for this purpose, we explore closely related alternatives that might be used in new approximate inference schemes.

1 Introduction

Much of modern statistics is conducted in the Bayesian setting, which suitably balances information from prior beliefs with that gained from new observations (data), all while allowing for sources of the relevant uncertainties to be communicated in a transparent fashion. In this setting, we are concerned with the observations $X$ and parameters $\Theta$. After making observations, we are typically interested in the posterior distribution of the parameters, $p(\Theta|X)$, which can be related to our prior beliefs in the parameters, the prior distribution $p(\Theta)$, the likelihood $p(X|\Theta)$ and the evidence $p(X)$ via Bayes’ rule

$$p(\Theta|X) = \frac{p(X|\Theta)p(\Theta)}{p(X)}.$$ (1.1)

∗Division of Systems and Control, Department of Information Technology, Uppsala university, SE-751 05 Uppsala, Sweden. jonas.evaeus@it.uu.se
This relation can, as shown by Bui-Thanh and Ghattas [2015], be viewed as a special case of an optimal solution to the problem of following the prior distribution while minimizing the error between model predictions and data in the mean squared error sense. Placing equal weight on each of these two terms in the optimization then precisely recovers Bayes’s rule.

For complex models, the form of the evidence means the relation in (1.1) is typically not tractable. Many of the challenges in Bayesian inference then amount to finding computationally viable methods for approximating the posterior distribution, and the development of such methods remains an active area of research. Two popular approaches (there exists many others!) will be presented in section 2, before section 3 concludes with a rather speculative sketch of possible avenues to explore for finding new efficient Bayesian inference methods.

2 Established Approaches to Bayesian Inference

One popular approach to Bayesian inference is to employ Markov Chain Monte Carlo Methods (MCMC), which collects samples from the posterior \( p(\Theta|X) \) via the stationary distribution of an ergodic Markov chain on \( \Theta \) which is equivalent to the posterior distribution itself. A subset of the collected samples then serve as an approximation of \( p(\Theta|X) \) (Bishop and Nasrabadi [2006]).

As outlined by Blei et al. [2017], the computational expense of MCMC methods can make them impractical in problems where data are plentiful or models overly complex. In such situations, a way around this hurdle is to use approximate techniques that allow for faster computation. One such technique is variational inference, which aims to choose from a family of candidate distributions, \( Q \), the distribution \( q^*(\Theta) \) that is in some sense closest to the actual posterior. While we will not delve into the details in this essay, the challenge then amounts to choosing a suitably inclusive family \( Q \) such that on the one hand, \( q^*(\Theta) \) is a sufficiently good approximation of the posterior, while on the other hand ensuring that the optimization is still feasible.

From what has been presented so far, one question that might naturally arise is: how do we measure the statistical difference between the distributions \( q(\Theta) \) and \( p(\Theta|X) \)? While there would appear to many candidates here, the seemingly arbitrary choice in the case of variational inference is the Kullback-Leibler (KL) divergence, and so the approximation of the posterior
is given by

\[
q^*(\Theta) = \arg \min_{q(\Theta) \in Q} KL(q(\Theta)||p(\Theta|X)),
\]  

(2.1)

where the KL divergence \(KL(q(\Theta)||p(\Theta|X))\) is given by

\[
KL(q(\Theta)||p(\Theta|X)) = \int q(\Theta) \log \left( \frac{q(\Theta)}{p(\Theta|X)} \right) d\Theta.
\]  

(2.2)

The KL divergence \(KL(p^0||p^1)\) was first introduced as a concept in information theory as a way to quantify the information gained in an observation that changes the probability distribution over possible outcomes from \(p^0\) to \(p^1\) (Kullback and Leibler [1951]). As would be expected from a statistical distance, it is non-negative and equal to zero only for the case of two identical distributions. As can be seen from even a quick glance at Equation (2.2), however, the KL divergence is not in fact symmetric w.r.t. the two distributions, and so is not a metric as defined by Kreyszig [1991] (in addition, it does not satisfy the triangle inequality). In practice, however, variational inference has proven to be incredibly useful for many setups, but also displays some shortcomings such as a tendency to underestimate the variance of the posterior (Blei et al. [2017]). The source of these shortcomings is not obvious, but to the author of this essay at least, it is conceptually unsatisfactory that the choice of measure of statistical difference used to set up the optimization problem does not appear to be the sole candidate. To illustrate this, we might instead consider minimizing the equally valid (but not equivalent!) divergence \(KL(p(\Theta|X)||q(\Theta))\). In general, this approach does indeed produce a different result, and is referred to in the literature as expectation propagation (Bishop and Nasrabadi [2006]). The issue is that the KL divergence quantifies the information gain when moving from one probability distribution to another, and so has a direction. It is therefore natural to consider the KL divergence when considering the information gain when updating prior beliefs to a posterior distribution, and this in fact forms a key part of the reasoning in Bui-Thanh and Ghattas [2015]. In the case of variational inference, however, we are merely seeking a distribution that is in some sense close to the true distribution, and so the asymmetric property of the KL divergence is unwarranted. The following section will consider other possible candidates for suitable measures of statistical difference that might be minimized in alternative approaches to approximate Bayesian inference.
3 Approximate Inference Beyond KL Divergence

Several conceptually satisfactory distances exist, and one example is the Hellinger distance which constitutes a metric as defined by Kreyszig [1991]. For two probability density functions $q(x)$ and $p(x)$, the square of the Hellinger distance is then given by

$$H^2(p, q) = \frac{1}{2} \int (\sqrt{p(x)}\sqrt{f(x)})^2 dx,$$  \hspace{1cm} (3.1)

which is clearly symmetric in $q$ and $p$ and satisfies $0 \leq H(p, q) \leq 1$. The issue with these candidates relative to the KL divergence is that they do not permit efficient optimization schemes (Blei et al. [2017]).

Before entirely giving up on the KL divergence as a valuable measure, it is worth noting that it has several desirable properties, as pointed out by Johnson and Sinanovic [2001], while importantly constituting a measure for which optimization is practically feasible. In a more general context than the specific problem of Bayesian inference considered here, the authors discuss the problems that arise from the asymmetry of the KL divergence. They present various avenues by which the KL divergence can be symmetrized to obtain a statistical difference measure with more desirable properties.

One such candidate is the $J$-divergence, which is simply the average of the two possible KL divergences, such that

$$J(p, q) = \frac{\text{KL}(p||q) + \text{KL}(q||p)}{2}.$$ \hspace{1cm} (3.2)

An algorithm approximating the posterior by minimizing $J(p(\Theta|X), q(\Theta))$ would then constitute a straightforward mix of the two approaches variational inference and expectation propagation considered in Section 2.

Johnson and Sinanovic [2001] furthermore present the resistor-average distance, which is constructed via the harmonic mean

$$\frac{1}{R(p, q)} = \frac{1}{\text{KL}(p||q)} + \frac{1}{\text{KL}(q||p)}.$$ \hspace{1cm} (3.3)

While there exist many other measures of statistical distance than those presented here (and it might be worth exploring those also!), the most obvious candidates to explore in the context of approximate Bayesian inference are those closely related to the KL divergences, which are already commonly used, as discussed in Section 2. For that reason, the candidates that are straightforward symmetrizations of the KL divergence appear to be a first
starting point for anyone looking to find a conceptually satisfying distance to minimize in an optimization scheme similar to that used in variational inference.

References


Solvability of integral equations on Lipschitz domains

David Krantz∗

June 23, 2022

Abstract

We apply the method of layer potentials to the Dirichlet problem for Laplace’s equation and investigate the solvability of the resulting integral equation \((\frac{1}{2}I + K)\varphi = g\) with regard to the regularity of the boundary \(\partial \Omega\) of the domain. When \(\partial \Omega\) is Lipschitz we prove the existence of a solution by showing that the operator \(\frac{1}{2}I + K\) is invertible on \(L^2(\partial \Omega)\), and briefly comment on the question of uniqueness. Lastly, we discuss some numerical challenges that arise when solving integral equations on Lipschitz domains.

1 Introduction

Partial differential equations can be found in various branches of mathematics and engineering. They can be used to describe complex natural phenomena such as electrodynamics, thermodynamics, fluid dynamics, and quantum mechanics. However, writing down explicit solutions to such equations is generally impossible. Therefore, it is highly desirable to research methods and develop efficient algorithms that can be used to numerically find approximate solutions to such equations. Doing this is unfortunately far from trivial as we in order to be successful must:

(I) Establish that the method yields a well-posed problem.

(II) Find a numerically feasible and stable algorithm for the chosen method.

∗Division of Numerical Analysis, Department of Mathematics, KTH Royal Institute of Technology, SE-100 44 Stockholm, Sweden. davkra@kth.se
The goal of this essay is to touch upon both of these problems in the setting of integral equations on Lipschitz domains.

Consider the (interior) Dirichlet problem for Laplace’s equation

\[
\Delta u = 0, \quad \text{in } \Omega \\
u = g, \quad \text{on } \partial \Omega,
\]  

(1.1)

where \( \Omega \subset \mathbb{R}^n, n \geq 3 \), is a bounded domain with connected boundary \( \partial \Omega \), and \( g \) is given boundary data. Due to the linearity of (1.1), it can be reformulated as an integral equation on the form,

\[(\lambda I + K)\varphi = g, \quad (1.2)\]

where the unknown \( \varphi \) and given data \( g \) are elements in some normed space \( X \), \( K : X \rightarrow X \) is an operator related to the Green’s functions or fundamental solution of the considered differential equation, \( I : X \rightarrow X \) is the identity operator, and \( \lambda \neq 0 \) is a real (or complex) number. Through the boundary integral method we approximate (1.2) in a finite dimensional space by the Nyström method, reducing it to a linear system of equations

\[Ax = b. \quad (1.3)\]

In Section 2, we give a preliminary invertibility result before we in Section 3 focus on (I) by studying the invertibility of the boundary integral operator \( \lambda I + K \) in terms of the regularity of \( \partial \Omega \). More specifically, we study the case when \( \Omega \) is a Lipschitz domain (see Verchota [1984] for a formal definition of a Lipschitz domain) and when \( K \) takes the form of a double-layer potential. Similar results to those presented in this section exists for the Neumann problem for Laplace’s equation (Verchota [1984]) and the Dirichlet problem for the Stokes equations (Fabes et al. [1988]) for various dimensions \( n \). Lastly, in Section 4 we connect to (II) by describing how (1.2) is reduced into (1.3), and comment on some of the numerical challenges that arise when solving this linear system of equations.

2 Preliminary invertibility result

The purpose of this section is to give a preliminary invertibility result that will form a structure for the proof of the main result presented in Theorem 3.1. More specifically, this preliminary result given in Lemma 2.2 relates the invertibility of a bounded linear operator to the associated dual spaces and adjoint operator. To show this, we first define the notion of an adjoint operator and then give a theorem useful for its proof.
**Definition 2.1** (Adjoint operator $T^\times$). Let $T : X \longrightarrow Y$ be a bounded linear operator, where $X$ and $Y$ are normed spaces. Then, the adjoint operator $T^\times : Y' \longrightarrow X'$ of $T$ is defined by

$$f(x) = (T^\times g)(x) = g(Tx), \quad g \in Y',$$

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

Using the notation in Definition 2.1 we give Theorem 2.1 from Yosida [1995], which is a part of the more commonly known closed range theorem proved by Banach [1932]. It provides a relation between the closedness a linear operator and its adjoint. In short, it says that $T$ has closed range if and only if $T^\times$ does.

**Theorem 2.1.** Let $X$, $Y$ be Banach spaces and $T : \mathcal{D}(T) \longrightarrow Y$ a closed linear operator such that $\mathcal{D}(T) = X$. Then, the following propositions are equivalent:

- $\mathcal{R}(T)$ is closed in $Y$.  
- $\mathcal{R}(T^\times)$ is closed in $X'$.

The proof of Theorem 2.1 presented in Yosida [1995] involves for example the Hahn-Banach theorem and the open mapping theorem, to which we refer the interested reader.

We are now ready to present Lemma 2.2.

**Lemma 2.2.** Let $H$ be a Hilbert space and $T : H \longrightarrow H$ be a bounded linear operator with the adjoint $T^\times : H' \longrightarrow H'$, where $H'$ is the dual space of $H$. Additionally, assume that the following holds:

- (i) $\mathcal{R}(T)$ is dense in $H$.
- (ii) $\mathcal{R}(T^\times)$ is closed in $H'$.
- (iii) $\mathcal{R}(T^\times)$ is dense in $H'$.

Then, $T$ is invertible on $H$.

**Proof.** In order to prove that $T$ is invertible, it suffices to show that it is bijective. We begin with proving that $T$ is surjective. From assumption (ii) we have that $T^\times$ has closed range in $H'$, which by Theorem 2.1 implies that $\mathcal{R}(T)$ is closed in $H$. Moreover, using assumption (i) we have that $\mathcal{R}(T)$ is dense in $H$, meaning that $T$ is surjective. Next, we show that $T$ is injective by proving that its null space $\mathcal{N}(T)$ is trivial. For a bounded linear operator
between two Hilbert spaces we have that $\mathcal{N}(W) = (\mathcal{R}(W^\times))^\perp$, which combined with assumption (iii) yields
\[ \mathcal{N}(T) = (\mathcal{R}(T^\times))^\perp = \left(\mathcal{R}(T^\times)^\perp\right)^\perp = H^\perp = \{0\}, \quad (2.2) \]
where we in the third equality used that for a linear subspace $X$ of a Hilbert space, $X^\perp = X^\perp$.

### 3 Invertibility of layer potentials

One way of characterizing the existence and uniqueness of the solution $\varphi$ to (1.2) is by the theorem often referred to as the Fredholm alternative, which relates it to the compactness of $K$. More specifically, it states that if $A$ is a compact linear operator, then (1.2) has a unique solution for every $g \in X$ if the corresponding homogeneous equation of (1.2) only has the trivial solution (Erwin [1978]). In order to relate this result to the regularity of $\partial \Omega$, we consider $K$ as a so-called boundary integral operator, given by
\[ (K\varphi)(x) = \int_{\partial \Omega} G(x,y)\varphi(y)dS_y, \quad x \in \partial \Omega, \quad (3.1) \]
where the kernel $G$ is defined for all $x, y \in \partial \Omega$, $x \neq y$, and possibly singular at $x = y$. Under the assumption that $G$ is a so-called weakly singular kernel and if $\partial \Omega$ is of class $C^1$, then the integral operator defined by (3.1) is compact as an operator on the Banach space $C(\partial \Omega)$ (Kress [2014]). Meaning that, in this case, the Fredholm theory can be directly applied in order to establish existence and uniqueness of $\varphi$. However, if $\partial \Omega$ is only Lipschitz the Fredholm theory is not readily applicable since we no longer can guarantee that $K$ will be compact. Hence, a different method is needed in order to determine the solvability of (1.2).

As previously mentioned, we study the case when $K$ takes the form of a double-layer potential, which we denote with $D$. When using this ansatz we seek solutions on the form $u(x) = (D\varphi)(x)$, resulting in $\lambda = 1/2$. This $1/2$ originates from the famous jump discontinuity that $D$ experiences as $x \in \Omega$ approaches a point $x_0 \in \partial \Omega$. It turns out that the compactness of the integral operator can be substituted by certain operator inequalities originating from so-called Rellich identities (Nečas [2012]). These inequalities are derived in Verchota [1984], where the author used them to prove the invertibility of $\frac{1}{2}I + D$ when $D$ is associated with a Lipschitz domain. We summarize this main invertibility result in the following theorem.
Theorem 3.1. Let $\Omega$ be a Lipschitz domain and the double-layer potential $D$ be defined as

$$(D\varphi)(x) = \int_{\partial\Omega} \frac{\partial \Psi(x,y)}{\partial \hat{n}} \varphi(y) dS_y, \quad x \in \partial \Omega,$$  \hspace{1cm} (3.2)

where $\Psi$ is the fundamental solution defined for all $x,y \in \partial \Omega$, $x \neq y$, of (1.1) (Evans [2010], Eq. 6) and $\hat{n}$ is a unit normal vector to $\partial \Omega$ pointing into $\Omega$. Then, $\frac{1}{2}I + D : L^2(\partial \Omega) \rightarrow L^2(\partial \Omega)$ is invertible.

Proof sketch. The proof consists of applying Lemma 2.2 to the operator $\frac{1}{2}I + D$ and proving that the conditions (i), (ii), and (iii) hold. Therefore, the proof is split into three parts. We now give a sketch of the proof for each step, and refer the reader to Verchota [1984] for the details:

(i) We prove that $R\left(\frac{1}{2}I + D\right)$ is dense in $L^2(\partial \Omega)$. If the operator $\frac{1}{2}I + D^\times$ is injective then $N(\frac{1}{2}I + D^\times) = \{0\}$. Using the relation $N(W) = (R(W))^\perp$ from the proof of (2.2) we get

$$\{0\} = N\left(\frac{1}{2}I + D^\times\right) = (R\left(\frac{1}{2}I + D\right)) \perp. \hspace{1cm} (3.3)$$

Furthermore, for a linear subspace $X$ of an inner product space, $(X^\perp)^\perp = X$ holds, which applied to (3.3) yields

$$R\left(\frac{1}{2}I + D\right) = \left(\left(R\left(\frac{1}{2}I + D\right)\right)^\perp\right) \perp = \{0\} \perp = L^2(\partial \Omega), \hspace{1cm} (3.4)$$

meaning that $R\left(\frac{1}{2}I + D\right)$ is dense in $L^2(\partial \Omega)$. Therefore, to prove condition (i) it suffices to show that $\frac{1}{2}I + D^\times$ is injective. The latter is proven by assuming $(\frac{1}{2}I + D^\times)f = 0$, and then showing that $f = 0$.

(ii) We prove that $R\left(\frac{1}{2}I + D^\times\right)$ is closed in $L^2(\partial \Omega)$. To do this, we show that there is a sequence $\{f_j\}_{j=1}^\infty$ such that $(\frac{1}{2}I + D^\times)f_j$ converges to some $g \in L^2(\partial \Omega)$ belonging to $R\left(\frac{1}{2}I + D^\times\right)$. By Lemma 4.8-7 in Erwin [1978], if the sequence $\{f_j\}_{L^2(\partial \Omega)}$ is bounded, then $f_j \rightharpoonup f$ weakly in $L^2(\partial \Omega)$. Thus, for any $h \in L^2(\partial \Omega)$

$$\int_{\partial \Omega} gh dx = \lim_{j \rightarrow \infty} \int_{\partial \Omega} \left(\frac{1}{2}I + D^\times\right) f_j h dx = \lim_{j \rightarrow \infty} \int_{\partial \Omega} f_j \left(\frac{1}{2}I + D\right) h dx$$

$$= \int_{\partial \Omega} f \left(\frac{1}{2}I + D\right) h dx = \int_{\partial \Omega} \left(\frac{1}{2}I + D^\times\right) fh dx. \hspace{1cm} (3.5)$$
Moreover, since $h$ was arbitrary, we get $g = (\frac{1}{2}I + D^x)f$, which clearly belongs to $\mathcal{R}(\frac{1}{2}I + D^x)$.

If $\|f_j\|_{L^2(\partial\Omega)}$ is not bounded we construct a normalized sequence $(\tilde{f}_j)$ as $\tilde{f}_j = f_j/\|f_j\|_{L^2(\partial\Omega)}$, resulting in $(\frac{1}{2}I + D^x)\tilde{f}_j \rightarrow 0$ and $\|\tilde{f}_j\|_{L^2(\partial\Omega)} = 1$. Then, by using similar arguments as above and utilizing that we in the previous part of this proof proved that $\frac{1}{2}I + D^x$ is injective in combination with the aforementioned operator inequalities based on Rellich identities, we show by contradiction that $(f_j)$ has to be bounded in $L^2(\partial\Omega)$. This then implies that $g$ belongs to $\mathcal{R}(\frac{1}{2}I + D^x)$.

(iii) We prove that $\mathcal{R}(\frac{1}{2}I + D^x)$ is dense in $L^2(\partial\Omega)$. We prove this by noting that it suffices to show that a dense subspace of $L^2(\partial\Omega)$ belongs to $\mathcal{R}(\frac{1}{2}I + D^x)$. To this end, we take $g \in C_0^\infty(\mathbb{R}^n)$ and show that its restriction to $\partial\Omega$ is contained in $\mathcal{R}(\frac{1}{2}I + D^x)$.

Theorem 3.1 and the known jump relations of the double-layer potential on Lipschitz domains implies the existence of a solution on the form of a double-layer potential to the Dirichlet problem. In order to establish uniqueness of the double-layer solution, additional conditions on the given data is needed. We present the following uniqueness result in the form of a simplified version of a theorem from Verchota [1984] without proof. Worth noting is that this concerns the uniqueness of solutions on the form of the double-layer potential. Thus, there may exist other solutions taking other forms.

**Theorem 3.2.** Let $\Omega \subset \mathbb{R}^n$, $n \geq 3$, be a connected domain with Lipschitz boundary $\partial\Omega$. Given boundary data $g \in L^2(\partial\Omega)$, then there exists a unique solution to the Dirichlet problem for Laplace’s equation (1.1). Moreover, the solution $u$ takes the form of the double-layer potential (3.2).

## 4 Discretization of integral equations on Lipschitz domains

We now comment on some of the numerical challenges of discretizing and solving integral equations on non-smooth boundaries, for example when $\partial\Omega$ is merely Lipschitz. The operator $K$ is not compact, its associated kernel $G$, and solutions $\varphi$ may be singular.

One common way of dealing with the first problem of compactness is to split the boundary into one part where it is smooth and another where it
is non-smooth. A point where $\partial \Omega$ is non-smooth is often referred to as a corner, corner point or singular point. For simplicity, assume that $\partial \Omega$ only has one corner and consider a boundary segment $\partial \Omega_\delta$ that excludes from $\partial \Omega$ a $\delta$-neighborhood of the corner point. Then, the integral operator

$$
(K_\delta \varphi)(x) = \int_{\partial \Omega_\delta} G(x, y) \varphi(y) dS_y, \quad x \in \partial \Omega_\delta,
$$

is compact, and the corresponding integral equation becomes

$$
(\lambda I + K_\delta) \varphi_\delta = g, \quad x \in \partial \Omega_\delta.
$$

Furthermore, the solutions $\varphi_\delta$ of (4.2) converge in a distributional sense to the solutions $\varphi$ of (1.2) (Bremer [2012]), meaning

$$
\int_{\partial \Omega_\delta} G(x, y) \varphi_\delta dS_y \rightarrow \int_{\partial \Omega} G(x, y) \varphi dS_y, \quad \text{as } \delta \rightarrow 0^+, \quad x \in \Omega^c.
$$

The problem of singular kernels and solution can be handled by refining the mesh towards the corner and by using various types of specialized quadrature rules. This refinement may however make the system of linear equations (of the form of (1.3)) ill-conditioned, resulting in a loss of accuracy. Worth mentioning is that this ill-conditioning is artificial in the sense that the approximations from the Nyström method converge in exact arithmetic. Although there exist methods that are able to alleviate this issue, it still is not a numerically feasible approach for large-scale domains with many corners as the linear system becomes exceedingly large. It is therefore of interest to develop compression techniques to reduce the dimension of the linear systems whilst retaining the accuracy of the solutions. One of these techniques that successfully has been applied to various types of problems is the scheme called ”recursively compressed inverse preconditioning” (Helsing and Ojala [2008]).

References

S. Banach. Théorie des opérations linéaires. 1932.


Weak Convergence and Its Properties

Gesina Menz*

June 17, 2022

Abstract

Strong convergence plays a big role in functional analysis as it lays the foundation for sets and spaces to be compact and as a foundation for many other areas of analysis. In comparison to strong convergence, a weaker concept called weak convergence exists. In this essay, we describe what weak convergence is, investigate how it can be used and what properties a weakly convergent set or space has by the Banach-Saks-Mazur Theorem and the Banach-Eberlein-Šmulian Theorem.

1 Introduction

To begin, we first need to define some concepts and background theory used in the Banach-Saks-Mazur Theorem or necessary for its proof.

The first important concept is strong convergence or as it is sometimes also called: norm convergence. As stated in Kreyszig [1978] it is defined as

Definition 1.1 (Strong convergence). A sequence \((x_n)\) in a normed space \(X\) is said to be strongly convergent (or convergent in the norm) if there is an \(x \in X\) such that

\[
\lim_{n \to \infty} \|x_n - x\| = 0.
\]

This is written

\[
\lim_{n \to \infty} x_n = x
\]

or simply

\[
x_n \longrightarrow x.
\]

*Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden. gesina.menz@it.uu.se
In comparison to strong convergence Kreyszig [1978] then defines weak convergence by

**Definition 1.2** (Weak convergence). A sequence \( (x_n) \) in a normed space \( X \) is said to be weakly convergent if there is an \( x \in X \) such that for every \( f \in X' \),

\[
\lim_{n \to \infty} f(x_n) = f(x).
\]

This is written as

\[ x_n \overset{w}{\to} \text{ or } x_n \rightharpoonup x. \]

The element \( x \) is called the weak limit of \( (x_n) \), and we say that \( (x_n) \) converges weakly to \( x \).

Definitions 1.1 and 1.2 show that strong convergence implies weak convergence but the converse of this is not necessarily true. So a strongly convergent sequence is always weakly convergent, however, a weakly convergent sequence is not always strongly convergent. In cases where strong convergence cannot be achieved, weak convergence can nevertheless be a useful tool to evaluate the behaviour of a sequence as will be shown in the following sections.

Additionally, we find the following theorem from Royden [2010]

**Theorem 1.1.** Let \( (x_n) \) be a weakly convergent sequence in the Hilbert space. Then \( (x_n) \) is bounded.

which we will use in the proof of the Banach-Saks-Mazur Theorem in Section 2.

## 2 Banach-Saks-Mazur Theorem

One of the main theorems concerning weak convergence is the Banach-Saks-Mazur Theorem which is also often called simply the Banach-Saks Theorem or the Banach-Saks property.

**Theorem 2.1** (Banach-Saks-Mazur Theorem). Let \( X \) be a Hilbert space and let \( (x_n) \) be a sequence in \( X \) such that

\[ x_n \rightharpoonup x \text{ as } n \to \infty \]

for \( x \in X \). Then there is a subsequence \( (x_{n_k}) \) of \( (x_n) \) for which
\[
\lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} x_{n_i} - x = \lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} x_{n_i} - x = 0,
\]

i.e. \( \frac{1}{k} \sum_{i=1}^{k} x_{n_i} \) converges strongly in \( X \).

**Proof.** Adapted from Royden [2010] and Gardner [2017].

Let \((x_n)\) be a sequence in a Hilbert space \( X \). By Theorem 1.1, there exists a subsequence \((x_{n_k})\) of \((x_n)\) which is bounded.

We can then set \( x_n = x_n - x \) with the assumption \( x = 0 \). Since the subsequence is bounded, we can choose \( M > 0 \) such that

\[
\|x_n\|^2 = \langle x_n, x_n \rangle \leq M \quad \forall n \in \mathbb{N}.
\]

Choosing \( n_1 = 1 \), we can then find \( n_2 \in \mathbb{N} \) such that \( |\langle x_{n_1}, x_{n_2} \rangle| \leq 1 \) since \( (x_n) \to x = 0 \) which implies \( \lim_{n \to \infty} \langle y, x_n \rangle = \langle y, 0 \rangle = 0 \quad \forall y \in X \). Thus, we see that

\[
\|x_{n_1} + x_{n_2}\|^2 = \langle x_{n_1} + x_{n_2}, x_{n_1} + x_{n_2} \rangle
= \langle x_{n_1}, x_{n_1} \rangle + 2\langle x_{n_1}, x_{n_2} \rangle + \langle x_{n_2}, x_{n_2} \rangle
= \|x_{n_1}\|^2 + \|x_{n_2}\|^2 + 2\langle x_{n_1}, x_{n_2} \rangle
\leq 2M + 2 \leq 2(2 + M).
\]

By induction, this shows that

\[
\|x_{n_1} + \cdots + x_{n_k}\|^2 \leq k(2 + M),
\]

so that

\[
\left\| \frac{x_{n_1} + \cdots + x_{n_k}}{k} \right\|^2 = \frac{1}{k} \sum_{i=1}^{k} x_{n_i} \leq \frac{2 + M}{k} \quad \forall k.
\]

Thus,

\[
\lim_{k \to \infty} \frac{1}{k} \sum_{i=1}^{k} x_{n_i} - x = \lim_{k \to \infty} \left\| \frac{1}{k} \sum_{i=1}^{k} x_{n_i} - x \right\| = 0.
\]

The Banach-Saks Theorem can also be extended to all uniformly convex normed spaces as shown in Kakutani [1939]. Hence, as long as they are uniformly convex, Banach spaces also have the Banach-Saks property.
3 Significance of Weak Convergence

As discussed in Narici and Beckenstein [2011], weak convergence was used extensively by Riesz and Banach, among others, during the emergence of functional analysis in the early 20th century alongside convergence in the norm. In this section, we will first introduce the Banach-Eberlein-Šmulian Theorem and then outline one way to use Theorem 2.1 together with the Banach-Eberlein-Šmulian Theorem to show why weak convergence and its resulting properties are useful tools.

3.1 Banach-Eberlein-Šmulian Theorem

The Banach-Eberlein-Šmulian Theorem is another important theorem concerning weak convergence. As we will need it alongside the Banach-Saks-Mazur Theorem later in this section, we will state it here. However, a proof will be omitted for brevity.

As stated in Ciarlet [2013], the Banach-Eberlein-Šmulian Theorem is

**Theorem 3.1** (Banach-Eberlein-Šmulian Theorem). (a) Any bounded sequence in a reflexive Banach space contains a weakly convergent subsequence.

(b) Conversely, a Banach space in which every bounded sequence contains a weakly convergent subsequence is reflexive.

A reflexive space, as defined in Kreyszig [1978], is a space $X$ for which the mapping from $X$ into its second algebraic dual space $X^{**}$ is surjective. Importantly, uniformly convex Banach spaces are reflexive so for all uniformly convex Banach spaces both Theorem 2.1 and 3.1 apply.

3.2 Weak Compactness

In finite dimensions any subset $M \subset X$, where $X$ is a normed space, is compact if and only if $M$ is closed and bounded. In infinite dimensions, however, a compact subset $M$ of a metric space must be closed and bounded but this is not a sufficient condition for compactness. As this simple way of classifying compactness is not generalisable to infinite dimensions, this increases the difficulty in determining if a set is compact.

If, however, a space $X$ for which the Banach-Saks-Mazur Theorem applies (i.e. Hilbert spaces and uniformly convex normed spaces) has a weakly convergent sequence this implies that the sequence is bounded by Theorem 1.1. Then by the Banach-Saks-Mazur and Banach-Eberlein-Šmulian Theorem any
subset of $X$ which is a weakly convergent sequence in $X$ is weakly closed, i.e. it is closed for the weak topology of $X$, and it shows weak compactness, as shown in Ciarlet [2013] and Narici and Beckenstein [2011].

In this way Theorem 3.1 implies that for example the closed unit ball $\tilde{B}$ of a Banach space is weakly compact if all its subsequences are weakly convergent.

Spaces with the Banach-Saks property, thus, keep some of the features of finite dimensional spaces as any subset is weakly compact if it consists of a weakly convergent subsequence which by itself implies boundedness. This results in spaces with the Banach-Saks property being easier to handle.

### 3.2.1 Example of a Weakly Compact Set

To illustrate a non-trivial weakly compact set which is not compact in the norm, we can consider the $c_0$ space. This is the space of all sequences $x = (\xi_j)$ of complex numbers converging to 0 with the metric induced by the $\ell^\infty$ space.

If we now take the set $X$ to be the set of the standard unit vectors and the zero vectors, i.e.

$$X : \{e_n\} \cup \{0\} \subseteq (c_0, \|\cdot\|_\infty),$$

this set is weakly compact by the Banach-Eberlein-Šmulian Theorem since any sequence $(x_n)$ in $X$ is a subsequence of $(e_n)$ with $x_n \rightarrow 0$. Now since the space $c_0$ is a space which has the Banach-Saks property, $X$ is weakly compact.

However, for the sequence $(e_n)$ we have that for all $i \neq j$

$$B_{1/2}(e_i) \cap B_{1/2}(e_j) = \emptyset,$$

i.e. the intersection of the open balls with radius $\frac{1}{2}$ around two non-identical unit vectors is empty, and, thus, the set does not contain any convergent subsequences. This in turn means that the set is not compact in the norm by definition (see e.g. Kreyszig [1978]).

### References


Nonlinear Control meets Functional Analysis

Anh Tung Nguyen*

June 20, 2022

Abstract

Due to the fact that over 99% of real-world systems are nonlinear, nonlinear control design has received much attention. This essay discusses the stability of nonlinear control systems through the lens of functional analysis. Some academic examples of nonlinear systems are given to demonstrate how to determine the stability of such systems.

1 Motivation

While we are manually operating a machine to help us finish a task, we want to go for Fika. What should we do? Should we stop the machine? The answer is “No”. We wish to employ some automatic control technique to automatically operate the machine and go for Fika all the time. Removing manual labor and automatically operating machines in our daily life are the main goals of developing automatic control technique. Thus, designing automatic control techniques is an interesting topic. Except for the economic concerns with operating systems, what is the most important criterion we need to achieve? The answer is stability. It can be briefly explained as follows: (Stability) i) the system operates to meet a given requirement; and then ii) the system never waives the requirement.

What are the majority of industrial controllers being used in automatically operating nonlinear systems? The nonlinear system models are normally linearized to become linear models. Then we will basically apply some famous linear control techniques to the linearized system model. Let us take a basic example to show how it works. Given a nonlinear system

\[ \dot{x}(t) = x^2(t) + u(t), \]  

(1.1)

*Division of Systems and Control, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden. anh.tung.nguyen@it.uu.se
where $x(t) \in \mathbb{R}$ is the system state and $u(t) \in \mathbb{R}$ is the system input.

**Step 1:** We need to find the equilibrium of the system (1.1) (see Def. 2.2). The equilibrium $x_e$ is defined as $\dot{x}_e(t) = 0$ and $u(t) = 0$, i.e., the system remains at $x_e$ forever without the external force $u(t)$.

$$0 = \dot{x}_e(t) = x_e(t)^2 + 0,$$  \hspace{1cm} (1.2)  
$$\Rightarrow x_e(t) = 0.$$  \hspace{1cm} (1.3)

This implies that $x_e$ is an equilibrium of the nonlinear system (1.1). Then, we hope to control the system stable at $x_e$ from any arbitrary initial state $x(0)$.

**Step 2:** We linearize the nonlinear model (1.1) by Taylor-series around the equilibrium $x_e$ with $\Delta x(t) = x(t) - x_e(t)$

$$\Delta \dot{x}(t) = 2x_e \Delta x(t) + \Delta u(t),$$  \hspace{1cm} (1.4)  
$$\Delta \dot{x}(t) = 0 + \Delta u(t).$$  \hspace{1cm} (1.5)

**Step 3:** We design a linear control input

$$\Delta u(t) = -\Delta x(t).$$  \hspace{1cm} (1.6)

Then, we have $\Delta x(t) \to 0$ or $x(t) \to x_e$. Let us show an illustration with $x(0) = 0.99$. However, if we choose $x(0) = 1.01$, we obtain $x(t) \to \infty$. This is a basic example to show that a linear controller might fail to meet our given requirement. This motivates us to study nonlinear control design to deal with nonlinear systems. Let us illustrate a very basic nonlinear controller to control the nonlinear system (1.1)

$$u(t) = -x^2(t) - x(t).$$  \hspace{1cm} (1.7)

Then, the system (1.1) is rewritten as follows

$$\dot{x}(t) = x^2(t) + u(t) = -x(t).$$  \hspace{1cm} (1.8)

We basically solve the above ODE to obtain $x(t) = x(0)e^{-t} \to x_e = 0$, $\forall x(0) \in \mathbb{R}$. Due to the fact that our example (1.1) is quite simple, we can easily solve the ODE after applying the nonlinear controller to confirm that $|x(t)| < \infty$ and $x(t) \to x_e$. However, when the system becomes highly nonlinear, it is no longer easily solved. Thus, we should skip finding analytic solution of ODEs. But, we still need to confirm that $|x(t)| < \infty$ and $x(t) \to x_e$. This essay handles such an issue via some functional analysis techniques without finding analytic solutions of ODEs.
2 Preliminaries

Throughout this essay, we will focus on the contraction mapping which is defined as follows

**Definition 2.1.** [Kreyszig, 1991, Def. 5.1-1] (Contraction) Let $X = (X, d)$ be a metric space. A mapping $T : X \to X$ is called a contraction on $X$ if there is a positive real number $\alpha < 1$ such that for all $x, y \in X$

$$d(Tx, Ty) \leq \alpha d(x, y). \quad (2.1)$$

Inspired by Def. 2.1, the following theorem gives us a result in complete spaces

**Theorem 2.1.** [Kreyszig, 1991, Th. 5.1-2] (Banach Fixed point theorem) Consider a metric space $X = (X, d)$, where $X \neq \emptyset$. Suppose that $X$ is complete and let $T : X \to X$ be a contraction on $X$. Then, $T$ has precisely one fixed point.

On the other hand, we also need to recall some definitions from control system theory

**Definition 2.2.** (Equilibrium points) [Khalil, 2015, Ch. 1] A point $X^*$ in the state space is said to be an equilibrium point of a non-autonomous nonlinear system (time-varying) $\dot{x} = f(x, t)$ if

$$x(t_0) = x^* \Rightarrow x(t) \equiv x^*, \forall t \geq t_0. \quad (2.2)$$

Further, for the autonomous nonlinear system $\dot{x} = f(x)$, the equilibrium points are the real solutions of the equation

$$f(x) = 0. \quad (2.3)$$

**Definition 2.3.** (Asymptotic stability of equilibrium points) [Khalil, 2015, Def. 3.1] Let the origin be an asymptotically stable equilibrium point of the system $\dot{x} = f(x) \in \mathbb{R}^n$, where $f$ is a locally Lipschitz function defined over a domain $D \subset \mathbb{R}^n$, $(0 \in D)$. The region of asymptotic stability is the set of all points $x_0 \in D$ such that the solution of

$$\dot{x} = f(x), \quad x(0) = x_0 \quad (2.4)$$

is defined for all $t \geq 0$ and converges to the origin as $t \to \infty$. Further, the origin is globally asymptotically stable if the region of asymptotic stability is the whole space $\mathbb{R}^n$. 

3 Nonlinear systems analysis

In this section, we will analyze and design controllers for examples of nonlinear systems via Th. 2.1.

3.1 Qualitative behavior near equilibrium points

To be more illustrative, we mainly focus on second order state-space model

\[
\begin{align*}
\dot{x}_1(t) &= f_1(x_1(t), x_2(t)) = x_2(t), \\
\dot{x}_2(t) &= f_2(x_1(t), x_2(t)) = -x_1(t)^3 - 0.2x_2(t).
\end{align*}
\]

(3.1)

Based on Def. 2.2, the origin \(x^* = [0, 0]^T\) is one of the equilibrium points of the system (3.1), i.e., \(f_1(0, 0) = 0\) and \(f_2(0, 0) = 0\). Next, let us show the phase portrait of the system (3.1) in Fig. 3.1a.

![Phase portrait](image)

Figure 3.1: a) Phase portraits of the system (1.1) with different initial states; b) a phase portrait with initial state \(x(0) = [-1.11, 0.34]^T\).

In Fig. 3.1a, we take three examples of initial states \(x(t = 0)\): 1) \(x(0) = [0.49, -0.24]^T\) associated with the blue trajectory; 2) \(x(0) = [0.79, 0.19]^T\) associated with the orange trajectory; and 3) \(x(0) = [-1.11, 0.34]^T\) associated with the green trajectory. We easily observe that all the trajectories eventually converge to the origin, the equilibrium point. However, we cannot verify the system (3.1) is asymptotically stable at the origin by choosing more initial states. Additionally, we are also not interested in solving the ODE representing the system (3.1). We need another method to ensure that the trajectories of all initial states in \(\mathbb{R}^2\) converge to the origin without explicitly finding solutions to the ODE (3.1).
3.2 Asymptotic stability analysis

In this section, we focus on analyzing the trajectory produced by choosing \( x(0) = [-1.11, 0.34]^\top \) (green trajectory in Fig. 3.1a). With respect to time \( t_1 = 1.25 \), let us construct a set

\[
V(x(t_1)) = \{ y = [y_1, y_2]^\top \in \mathbb{R}^2 \mid y_1^4 + 2y_2^2 = x_1(t_1)^4 + 2x_2(t_1)^2 \}. \tag{3.2}
\]

Analogously, we also construct a similar set at time \( t_2 = 6.25 \). Those sets are illustrated in Fig. 3.1b, i.e., purple line for \( V(x(t_1 = 1.25)) \) and red line for \( V(x(t_1 = 6.25)) \).

Intuitively, the set \( V(x(t)) \) is a contraction mapping as time goes to \( \infty \). On the other hand, with the function \( V(x(t)) = x_1(t)^4 + 2x_2(t)^2 \), we can also show that

\[
\dot{V}(x(t)) = \frac{\partial V(x(t))}{\partial x} \frac{dx}{dt} = -0.8x_2^3 \leq 0, \tag{3.3}
\]

\[
\dot{V}(x(t)) = 0 \iff x_2(t) \equiv 0, \quad x_1(t) = 0, \tag{3.4}
\]

\[
\Rightarrow \begin{cases} 
V(x(t_2)) < V(x(t_1)), \quad \forall t_2 > t_1, \\
V(x(t_2)) = V(x(t_1)) \iff x(t_2) = x(t_1) = [0, 0]^\top. 
\end{cases} \tag{3.5}
\]

Based on Th. 2.1, (3.5) implies that \( V(x(t)) \) is a contraction mapping and the origin is a fixed point. Therefore, the system (3.1) is globally asymptotically stable at the origin (see Def. 2.3).

3.3 Nonlinear controller design

Assume that the nonlinear system (3.1) is modified as follows

\[
\begin{align*}
\dot{x}_1(t) &= x_2(t), \\
\dot{x}_2(t) &= -x_1(t)^3 + u(t),
\end{align*} \tag{3.6}
\]

where \( u(t) \in \mathbb{R} \) is control input. Without input \( u(t) = 0 \), the phase portrait of the system (3.6) is illustrated in Fig. 3.2a. We simply observe that the system does not converge to the origin. Thus, we need to design \( u(t) \) to ensure that the system converges asymptotically to the origin. We basically select \( u(t) = -0.2x_2(t) \) to convert the system (3.6) into (3.1) which is already said to be asymptotically stable at the origin. However, this is not the only control input that asymptotically stabilize the system. Let us choose a different control input as follows

\[
u(t) = -x_1(t) - x_2(t) - 3x_1(t)x_2(t)(x_1(t) + x_2(t)) - x_2(t)^3, \tag{3.7}\]

\]
and the following function

\[ V(x(t)) = \frac{1}{2} x_1(t)^2 + \frac{1}{2} (x_1(t) + x_2(t))^2. \]  

(3.8)

Taking time-derivative of function \( V(x(t)) \) in (3.8), one has

\[ \dot{V}(x(t)) = -x_1^2 - (x_1(t) + x_2(t))^4 < 0, \]  

(3.9)

\[ V(x(t_2)) < V(x(t_1)), \quad \forall \, t_2 > t_1. \]  

(3.10)

In light of Th. 2.1, there exists a fixed point which is the origin since \( V(x(t)) = 0 \Leftrightarrow x(t) = [0, 0]^T \). This also confirms that the designed nonlinear control input (3.7) is able to asymptotically stabilize the nonlinear system (3.6). An illustration of some example trajectories is given in Fig. 3.2b. From any example initial states, the control input (3.7) is able to drive the system (3.6) to the origin.

4 Conclusion and open questions

In conclusion, the result of Banach fixed point theorem (Th. 2.1) gives us a powerful tool to analyse the stability of nonlinear systems. By leveraging such stable nonlinear systems, we simply design controllers for unstable nonlinear systems such that they are twins of the stable systems. Additionally, we also design controllers and function \( V(x(t)) \) e.g., (3.7) and an associated function (3.8) that ensure the asymptotically stability of nonlinear systems.
Due to the fact that we considered a very simple nonlinear system, we simply find a linear or nonlinear control input (3.7) and function $V(x(t))$ as (3.8). However, real-world systems are represented by very complex nonlinear functions. It is challenging to find a proper function that answers whether the system is asymptotically stable at the equilibrium points. To the best of our knowledge, there is no concrete procedure to find such functions for general nonlinear systems.

**References**


Stability analysis with the Koopman and Perron-Frobenius operators

Håkan Runvik∗

June 11, 2022

Abstract

This essay discusses how stability, which is an important aspect in dynamical systems and automatic control, can be analyzed using tools from functional analysis. Two adjoint linear operators are considered: the Koopman operator, which describes the time evolution of functions of the system state, and the Perron-Frobenius operator, which operates on densities.

1 Introduction

We will investigate how two dual operator-theoretic frameworks can be used for stability analysis of dynamical systems. The first is based on the Koopman operator, which describes the time evolution of functions of the system state. Originally developed in the 1930s to describe ergodic properties of measure-preserving systems, the Koopman operator has recently been applied to a wider class of dynamical systems. See e.g. Brunton et al. [2022] for an overview of the Koopman operator theory. Concretely, results from Mauroy and Mezić [2016] will be reviewed.

The second framework is based on densities, using the Perron-Frobenius (P-F) operator. Here the stability results will come from Vaidya and Mehta [2008]. As noted in Mauroy and Mezić [2016], it is surprising that even though these operators have been known for a long time, and Lyaponov’s second method for stability (see Section 2.1) implicitly uses an operator-theoretic framework, explicit operator-theoretic stability analysis is a subject that only recently has received attention in research.

∗Division of Systems and Control, Department of Information Technology, Uppsala university, SE-751 05 Uppsala, Sweden. hakan.runvik@it.uu.se
1.1 The Koopman and P-F operators

Consider the discrete dynamical system
\[ x_{k+1} = T(x_k), \quad (1.1) \]
where \( x_k \in S \subseteq \mathbb{R}^n \) and \( T : S \rightarrow S \) and let \( \mathcal{F} \subseteq C^0(S) \) be a space of continuous functions \( f : S \rightarrow \mathbb{R} \), denoted as observables of the system. The Koopman operator \( U : \mathcal{F} \rightarrow \mathcal{F} \) is then defined by the composition
\[ Uf(x) = f(Tx). \]

Much of the utility of the Koopman operator comes from its linearity; if \( f_1, f_2 \) are two observables, we get
\[ U[f_1 + f_2](x) = [f_1 + f_2](T(x)) = f_1(T(x)) + f_2(T(x)) = Uf_1(x) + Uf_2(x). \]

On the other hand, the main disadvantage is that the operator is infinite-dimensional.

If the space \( \mathcal{F} \) is equipped with the inner product
\[ \langle f, \rho \rangle = \int_S f(x)\rho(x) \, dx, \]
we define the P-F operator \( P : \mathcal{F} \rightarrow \mathcal{F} \) as the adjoint of the Koopman operator, i.e.
\[ \langle Uf, \rho \rangle = \langle f, P\rho \rangle. \]

If \( T \) is bijective and differentiable with Jacobian \( J \), the variable substitution \( y = T^{-1}x \) gives
\[ \langle Uf, \rho \rangle = \int_S f(T(x))\rho(x) \, dx = \int_S f(y)\rho(T^{-1}(y))/|J(y)| \, dy, \]
so the P-F operator is given by
\[ P\rho(x) = \frac{\rho(T^{-1}(x))}{|J(x)|}. \]

In the P-F framework, \( \rho \) is normally the density of a measure \( \mu \), i.e. \( \mu(A) = \int_A \rho(x) \, dx \), which enables a more general definition of the P-F operator. We then let \( \mathcal{B}(S) \) be a Borel \( \sigma \)-algebra and \( \mathcal{M}(S) \) be the vector space of real-valued measures on \( \mathcal{B}(S) \), and define the P-F operator \( \mathbb{P} : \mathcal{M}(S) \rightarrow \mathcal{M}(S) \),
\[ \mathbb{P}[\mu](A) = \mu(T^{-1}A), \]
where \( A \) is a set in \( \mathcal{B}(S) \) and \( \mu \) is a measure in \( \mathcal{M}(S) \). See Vaidya and Mehta [2008] for further details.
1.1.1 Continuous time systems

The Koopman and P-F operators also apply to continuous system of the form

\[ \dot{x} = F(x), \quad x \in S \subseteq \mathbb{R}_n, \]  

(1.2)

with the corresponding flow map \( \varphi^t \). A key difference compared to the discrete time case is that rather than a single Koopman operator, a one-parametric semi-group of operators \( \{U_t\}_{t \geq 0} \), with elements

\[ U_t f(x) = f(\varphi^t) \]

is used. Since this case is similar, but more involved compared to the discrete time case, we will not cover it further here, but instead refer to the description in, e.g., Brunton et al. [2022] and Rajaram et al. [2010].

2 Stability analysis

We now review some basic stability analysis of systems of the form (1.1).

In the linear time-invariant case, \( T(x) = Ax \), where \( A \) is an \( n \times n \) matrix. This system is asymptotically stable, if for all initial conditions, the state converges to the fixed point \( x = 0 \) as \( k \to \infty \). The stability is determined by the eigenvalues \( \lambda_i \) of \( A \), the system is asymptotically stable if and only if all eigenvalues satisfy \( |\lambda_i| < 1 \). The corresponding condition for stability of a linear continuous time system is \( \text{Re}(\lambda_i) < 0 \).

In the nonlinear case, linearization around a fixed point provides local stability conditions. A fixed point \( x^* \) is then asymptotically stable, if there exists a \( \delta \) such that for all \( x_0 : |x_0 - x^*| < \delta \), \( T^k(x_0) \to x^* \) as \( k \to \infty \), where \( T^k \) denotes applying \( T \) \( k \) times. This stability is established from the eigenvalues of the Jacobian matrix in an analogous way to the linear case. It is also of interest, but generally harder, to establish global stability for a nonlinear system. in this case, an operator-theoretic approach can be used.

2.1 Lyapunov’s second method for stability

Lyapunov’s second method is often used to establish stability of nonlinear dynamical systems. For the discrete time system (1.1) with a fixed point \( x^* \), it utilizes a Lyapunov function \( V : S \to \mathbb{R} \), which has the properties

\[ V(x) \geq 0, \quad V(x) = 0 \iff x = x^* \]  

(2.1)

\[ V(T(x)) < V(x) \quad \forall x \in S - \{x^*\} \]  

(2.2)

\( ^1 \)The mapping from an initial state to the state at time \( t \) for system (1.2).
If a Lyapunov function exists, $x^*$ is globally asymptotically stable. For continuous time systems, the Lyapunov function is defined in an analogous fashion to get a corresponding stability result. Since $V$ is an observable of the system, and the property (2.2) is defined by the action of the Koopman operator, this is in fact an operator-based method.

2.2 Stability analysis using the eigenfunctions of the Koopman operator

Some results from Mauroy and Mezić [2016] will be reviewed here, so references to Propositions, Theorems, etc. in this section will refer to that work. The global stability of attractors (a generalization of a fixed point to sets, see Definition 2) and hyperbolic fixed points of continuous time dynamical systems of the form (1.2) is considered. Much of the analysis is based on spectral properties of the Koopman operator, see Chapter 7 in Kreyszig [1991].

2.2.1 Decomposition of the Koopman operator

For a given attractor $A$, a subspace $F_{A_c} \subseteq F$ is introduced, containing all functions with support on the complement $A_c$ of the attractor. A necessary and sufficient condition for the attractor to be globally attractive in $S$ is given in Proposition 1 as

$$\lim_{t \to \infty} U_t^{A_c} f = 0 \quad \forall f \in F_{A_c},$$

where $U_t^{A_c}$ is the restriction of the Koopman operator to $A_c$.

2.2.2 Koopman eigenfunctions

Proposition 1 has limited utility since it considers all functions in $F_{A_c}$. An approach based on the eigenfunctions of the Koopman operator is therefore developed. The main result is Theorem 1, which states that if $\phi_\lambda(x)$ is an eigenfunction corresponding to the eigenvalue $\lambda$ with $\text{Re}(\lambda) < 0$, the zero level set

$$M_0 = \{ x \in S\phi_\lambda(x) = 0 \}$$

is forward invariant\(^2\) and globally asymptotically stable. By Corollary 1, the intersection of such sets is also globally asymptotically stable. In the typical case, global stability analysis of an attractor $A$ can then be performed using a finite number of eigenfunctions, related to the dimension of the attractor and the system.

\(^2\)For all initial conditions in $M_0$, the state of the system will remain in $M_0$.\n
In Proposition 2, the result is specialized to a hyperbolic fixed point $x^*$. $x^*$ is globally stable if and only if the Koopman operator has $n$ eigenvalues $\lambda_i$ with $\text{Re}(\lambda_i) < 0$ and $\nabla \phi_{\lambda_i}(x^*) \neq 0$. A formula for creating a Lyapunov function (see Section 2.1 above) from these eigenfunctions is also provided. In both Theorem 1 and Proposition 2, the requirement $\text{Re}(\lambda_i) < 0$ shows the connection between the Koopman approach and the stability analysis in the linear or linearized case.

2.2.3 Numerical methods

To utilize Proposition 1 and 2 to evaluate the stability of a system, Koopman eigenfunctions need to be computed. Numerical methods are generally required to do this; in the paper two polynomial-based methods are presented for this purpose. The first is based on Taylor series. It can be used to get a conservative estimate of the basin of attraction, but only when the eigenfunctions are analytic. The second uses Bernstein polynomials, and has the advantage of good convergence even for non-analytic eigenfunctions. A downside of this method is that the results are inaccurate in some cases, such as if $S$ contains an unstable fixed point. The methods can therefore be seen as complementing each other.

2.3 Lyapunov measure for almost everywhere stability

A brief overview of results from Vaidya and Mehta [2008] is given here. Similarly to Mauroy and Mezić [2016], the stability analysis utilizes an operator decomposition based on an attractor set $A$. Here, the P-F operator $\mathbb{P}$ is decomposed into $\mathbb{P}_0$ and $\mathbb{P}_1$, where $\mathbb{P}_0$ operates on measures restricted to $A$ and $\mathbb{P}_1$ is restricted to $A_c$. For $\mathbb{P}_0$, invariant measures are of interest, i.e. measures $\mu$ such that

$$P[\mu](A) = \mu(A).$$

This can also be expressed as $\mu$ being a fixed point of $\mathbb{P}$, or that $\mu$ is an eigenmeasure corresponding to the eigenvalue 1. Invariance captures the recurrence of an attractor set, i.e. that the system will return to states arbitrarily close to initial states in the attractor.

For the operator $\mathbb{P}_1$, a Lyapunov measure $\mu$, defined by the property

$$\mathbb{P}_1 \mu < \alpha \mu,$$  \hspace{1cm} (2.3)

where $\alpha < 1$, is sought. Theorem 17 states that the existence of such a measure implies the stability of the set $A$ in an almost everywhere sense. A rough interpretation of this theorem and its proof is that for the system to
end up in state in $A_c$, it has to start from an initial condition which has measure zero. This is a result of the shrinkage of the Lyapunov measure given by (2.3). As a result, for almost all initial conditions, the system will end up in $A$.

For numerical computations, a finite partitioning of the phase-space is suggested. Theorem 28 shows that with this approximation, a Lyapunov measure can be approximated (if it exists) through solving a system of linear inequalities. Linear programming can thus be used to solve the problem.

3 Discussion

In this work, two operator-theoretic frameworks have been presented, and results from two publications which use these operators for stability analysis have been reviewed. Stepping back from this particular topic to modeling more broadly, the existing literature indicates that the Koopman approach is significantly more popular than methods based on the P-F operator. Presumably, the Koopman operator lends itself better to computations and numerical methods, e.g., the discussion in Remark 1 in Mauroy and Mezić [2016] indicates that this could be the case. However, a further investigation of this is a topic for another time.

References


Sobolev inner-product spaces with applications to finite element methods

Ivy Weber *

June 10, 2022

Abstract

Sobolev inner-product spaces, or Hilbert-Sobolev spaces, are used to define function spaces for the finite element method (FEM) in which to search for numerical weak solutions to partial differential equations. This mini-essay discusses some of the functional analysis related to these function spaces, including properties such as completeness, and provides an overview of how they are used in FEM. Finally a brief discussion of different spaces and their potential to improve stability of time-stepping solvers Weber et al. [2022] is made.

1 Introduction

When using the finite element method (FEM) Larson and Bengzon [2013], choosing appropriate function spaces for the method has a significant effect on the validity and accuracy of the method. There are two points at which a space must be chosen: first a decision about which function spaces the solution and test functions should theoretically lie within, and then a decision about which function spaces to work with computationally.

The typical choice for the first decision is to use a Sobolev inner-product space, or Hilbert-Sobolev space, for all functions. These were developed by Sergei L’vovich Sobolev as a way to study weak solutions to differential equations Sobolev [1963a] Sobolev [1963b].

The second choice is necessary due to the nature of modern computing, since it is not possible to model an infinite-dimensional function space such as a Sobolev space directly. It is therefore necessary to choose a finite basis to

*Division of Scientific Computing, Department of Information Technology, Uppsala University, SE-751 05 Uppsala, Sweden. ivy.weber@it.uu.se
approximate the space as finite-dimensional subspace of the original. A lot of the ongoing research about FEM, as well as other numerical methods such as radial basis functions, concern the choice of this basis. The aim of this essay is to provide a brief introduction into the motivation behind the use of Hilbert-Sobolev spaces and the method of choosing a finite basis used by the finite element method. The motivation for this is the importance of both parts to fully understanding the finite element method. For this reason the discussion begins with some background required for Sobolev spaces and a brief overview of FEM, before introducing the definition of Sobolev spaces and later their use in FEM. Finally a brief discussion of how a slight modification to the usual Hilbert-Sobolev space $H^1$ can bring certain numerical benefits is made, in order to emphasise that an improved understanding of the underlying functional analysis is still useful on a practical level.

For the purposes of this mini-essay discussion is limited to the case for 1D, however all the concepts discussed here generalise well to higher dimensions.

## 2 Preliminary material

### 2.1 Finite element methods

Finite element methods are a class of numerical solution methods for PDEs $^1$. The general approach for linear PDEs $^2$ can be summarised as follows:

1. Discretise the spatial domain of the equation into a finite number of cells or *elements*,

2. Define test and trial spaces of functions over the discretised domain. Typically the same function space will be used for both to obtain useful matrix properties, and this space is often a Hilbert-Sobolev space (see section 3.1),

3. Construct the weak form of the PDE by taking inner products with an arbitrary function from the test space and making the ansatz that the solution is in the trial space,

---

$^1$partial differential equations  
$^2$Nonlinear equations generally require some extra care to produce stable numerical methods, and there is no one simple approach that always works. The discussion is restricted to linear equations for simplicity
4. Construct a finite set of basis functions with compact support \(^3\) to approximate the test and trial spaces,

5. Express weak form of the equation as either a finite-dimensional linear equation or time-dependent ODE \(^4\) system. \(^5\)

Aside from the choice of domain elements the most significant factor to the accuracy and stability of these methods is the choice of function spaces and the finite-dimensional approximations of these spaces. The weak form of the equation means that Hilbert-Sobolev spaces \(^3.1\) are the most fitting candidates.

### 2.2 Lebesgue integrals

A brief discussion of what is meant by integrability is needed to define Sobolev spaces \(^6\). Typically the type of integral of interest is a Lebesgue integral Florescu [2021]. This approach discretises the range of the function rather than the domain, then takes the limit of sums of the measures of certain subsets:

1. Divide the range of values \([y_{\text{min}}, y_{\text{max}}]\) that the function \(f(x)\) takes into a sequence of sub-intervals given by \(y_{\text{min}} = y_0 < y_1 < \cdots < y_N = y_{\text{max}}\),

2. Evaluate the sets \(E_i = \{x : y_{i-1} \leq f(x) < y_i\}\), \(i = 1, \ldots, N - 1\) and \(E_N = \{x : y_{N-1} \leq f(x) \leq y_{\text{max}}\}\) \(^7\),

3. For some measure \(\lambda(E)\) \(^8\), define the sums

   • \(\sigma_N = \sum_{j=1}^{N} \lambda(E_j) y_{j-1}\),

   • \(\Sigma_N = \sum_{j=1}^{N} \lambda(E_j) y_j\),

4. If both sums converge to the same value, then the function is Lebesgue integrable and the limiting value is the value of the integral.

---

\(^3\)Compact support means that there is only a small region of the spatial domain where the function has a non-zero value

\(^4\)ordinary differential equation

\(^5\)This is the step where linearity of the equation simplifies the process, since the system can be constructed directly by substituting in basis functions.

\(^6\)Or any space that can be used with FEM, since the weak integral form of the equation is used.

\(^7\)\(E_N\) is defined separately to remove the strict inequality. This is necessary in case the set of values of \(x\) where \(f(x) = y_{\text{max}}\) has a non-zero measure.

\(^8\)If these sets are not measurable, then the function is not Lebesgue integrable.
A full discussion of measure theory and what is meant by measurability is beyond the scope of this essay, but the idea is to generalise the concept of interval lengths to more arbitrary sets. This approach is more general than the Riemann integral and also allows more functions to be integrated, such as the function \( I_Q \) on \([0, 1]\):

\[
I_Q(x) = \begin{cases} 
1 & x \in \mathbb{Q} \cap [0, 1], \\
0 & x \in (\mathbb{R} \setminus \mathbb{Q}) \cap [0, 1], 
\end{cases}
\]

which is Lebesgue integrable but not Riemann integrable.

A couple of important results for Lebesgue integration are as follows:

- A function that is Riemann integrable is Lebesgue integrable, and the integrals give the same result (provided a metric consistent with interval length is used),
- A function that is continuous almost everywhere is Lebesgue integrable.

The Lebesgue integral can be used to define normed spaces, denoted \( L^p \), of functions which are Lebesgue integrable. Since these are one instance of the more general Sobolev spaces, a definition is left until section 3.

### 2.3 Weak derivatives

In many applications it might not be practical to meet the criteria for a conventional (strong) derivative of a given order to exist across the entire domain. However, it is often possible to define a function that behaves similarly enough to a strong derivative to be useful. A function \( v(x) \) on the interval \( x \in [a, b] \) is called a weak derivative of another function \( u(x) \) if it satisfies

\[
\int_a^b v(x)\phi(x)dx = -\int_a^b u(x)\phi'(x)dx,
\]

for every \( \phi \in C_0^\infty[a, b] \) \(^9\). This concept can be extended to arbitrarily high order derivatives to define a weak \( n^{th} \)-order derivative \( v_n(x) \):

\[
\int_a^b v_n(x)\phi(x)dx = (-1)^n \int_a^b u(x)\phi^{(n)}(x)dx,
\]

where \( \phi^{(n)} \) is the conventional \( n^{th} \) derivative of \( \phi \). For the sake of clarity, for the rest of this essay these weak derivatives are denoted by \( v_n(x) = \partial_n u(x) \). There are several useful results regarding weak derivatives. The main two are as follows:

\(^9\)The space of infinitely differentiable functions on \([a, b]\) with compact support
Lemma 2.1. The weak derivative can be seen as an extension of the strong derivative due to the following properties:

1. If two functions \( v_1 \) and \( v_2 \) are weak derivatives of function \( u \) then the set of values where they differ has measure 0,

2. If a function \( u \) has a strong derivative \( w \), then \( w \) is a weak derivative of \( u \).

Proof. 1. If \( v_1, v_2 \) are weak derivatives of the same function, then for any \( \phi \in C^\infty_{\text{loc}}[a,b] \)

\[
\int_a^b v_1(x)\phi(x)dx = \int_a^b v_2(x)\phi(x)dx,
\]

\[
\implies \int_a^b (v_1(x) - v_2(x)) \phi(x)dx = 0,
\]

which can only be possible if the integral of \( v_1 - v_2 \) over every subinterval \([c,d]\) with \( a < c \leq d < b \) is 0.

2. Let \( \frac{du}{dx} = w \). For every \( \phi \in C^\infty_{\text{loc}}[a,b] \) the definition gives \( \phi(a) = \phi(b) = 0 \). Using integration by parts gives

\[
\int_a^b w(x)\phi(x)dx = [u(x)\phi(x)]_a^b - \int_a^b u(x)\phi'(x)dx = -\int_a^b u(x)\phi'(x)dx.
\]

An important consequence of 1 is that weak derivatives are unique only up to a quotient class of functions.

3 Sobolev normed spaces

To use weak derivatives in a practical setting, it’s important to know the functions being used actually have weak derivatives that can be used. Sobolev spaces Nikol’skii (usually denoted \( W^{r,q} \), or \( W^{\alpha,q} \) in higher dimensions) are vectors spaces of functions for which weak derivatives up to order \( r \) (or order array \( \alpha \) in higher dimensions) exist, and lie in the Lebesgue space \( L^q \).

The fact that Sobolev spaces are vector spaces follows immediately from the definition of a weak derivative. They can be made into normed spaces in the following way, where \( q \geq 1 \) is a parameter used to define the norm:

\[
||u||_{W^{r,q}} = \left\{ \begin{array}{ll}
\left[ \sum_{j=0}^r \int_a^b |\partial_j u(x)|^q dx \right]^{\frac{1}{q}}, & 1 \leq q < \infty, \\
\sum_{j=0}^r \text{ess sup}_{x \in [a,b]} |\partial_j u(x)|, & q = \infty,
\end{array} \right.
\]  

(3.1)
where the essential supremum is defined as follows
\[
\text{ess sup}_{x \in [a, b]} [f] = \inf \left\{ \theta : \int_a^b I[f(x) > \theta] \, dx = 0 \right\},
\]
\[
I[X] = \begin{cases} 
1 & X \text{ is true}, \\
0 & X \text{ is false}. 
\end{cases}
\]

The essential supremum can be thought of as the largest value that a function approaches in a domain over a non-negligible subset (the Lebesgue measure of that subset is non-zero).

The above functions clearly satisfy the positivity and scalar multiplication criteria for a norm. The triangle inequality follows from the Minkowski inequality for functions, a proof of which is given in Florescu [2021]. Uniqueness is solved by identifying the members of a Sobolev space as quotient classes of functions rather than individual functions. For instance in the spaces \( W^{0, q} = L^q \) any two functions are equivalent if the measure of their difference is 0 (they are equivalent almost everywhere). Without doing this, the above functions would define semi-norms.

Sobolev spaces are also Banach, or complete. This follows from the convergence properties of the Lebesgue integral and completeness of the Lebesgue spaces Florescu [2021].

### 3.1 Inner product spaces \((q = 2)\)

In the special case of a Sobolev space with \( q = 2 \) we denote the space \( W^{1,2}[a, b] \) by \( H^r[a, b] \), and the norm (3.1) is induced by the inner product
\[
\langle u, v \rangle_{H^r} = \sum_{j=0}^{r} \left[ \int_a^b \partial_j u(x) \partial_j v(x) \, dx \right].
\] (3.2)

Using the earlier result that \( W^{r,q} \) is a Banach space, this immediately shows that \( H^r \) is a Hilbert space. This motivates the label ‘Hilbert-Sobolev space’ for any Sobolev space with \( q = 2 \), and allows the theory developed for complete inner product spaces to be applied to these functions.

### 4 Use in finite element methods

As stated in the introduction finite element methods proceed by rewriting a partial differential equation (the strong form of the problem) as a weak
integral equation (the weak form). An example of this process is briefly outlined below for the Poisson equation in 1D on a domain \([a, b]\):

\[
\begin{cases}
-u''(x) = f(x) & x \in (a, b), \\
u(a) = g_1, \\
u(b) = g_2,
\end{cases}
\]

Integrate with some \(v(x)\):

\[-\int_a^b u''(x)v(x)\,dx = \int_a^b f(x)v(x)\,dx,
\]

Integrate by parts:

\[\int_a^b u'(x)v'(x)\,dx - [v(x)u'(x)]_a^b = \int_a^b f(x)v(x)\,dx.
\]

This expression indicates that the main requirement on both \(u\) and \(v\) is the existence of all integrals. This can be satisfied by replacing the strong derivatives in the above expression with \(\partial u, \partial v\) and requiring \(u, v \in H^1[a, b]\), as well as \(f \in L^2[a, b]\).

It is also possible to define a subspace \(H^1_0[a, b] \subset H^1[a, b]\) of functions that are uniformly zero at \(a\) and \(b\) to eliminate the boundary contributions. The solution can then be approximated as \(u = u_0 + g_u\), where \(u_0 \in H^1_0[a, b]\) and \(g_u\) is some function satisfying \(g_u(a) = g_1, g_u(b) = g_2\) that becomes 0 a small finite distance into the domain.

The space \(H^1[a, b]\) is generally infinite dimensional, so it is necessary to choose a finite dimensional subspace to work with. The usual approach uses the fact that polynomials are dense in the space of continuous functions (which is usually another desirable property for a solution), so for each element (see section 2.1) a polynomial basis up to a fixed degree \(p\) is defined. These polynomials are then extended to the entire domain by matching values at the element boundaries with other piecewise polynomials. When done correctly, this results in a finite set of continuous functions (continuity is sufficient for a function to lie in \(H^1[a, b]\)) for which each member only has a small region of non-zero support. This makes it possible to use part of the basis to approximate \(H^1_0[a, b]\) and the remaining part to approximate \(g_u\). The span of this finite basis will be denoted \(DH^1[a, b]\) to avoid confusion with the full infinite dimensional \(H^1[a, b]\).

The order of accuracy can be raised in two different ways. One way is by adjusting the value of \(p\). The set of all polynomial functions is dense subset of \(H^1[a, b]\), which means that under exact arithmetic FEM can obtain an arbitrarily high-order accuracy in terms of the grid size \(h\) by increasing it. This approach often comes with certain computational costs, such as reducing the sparsity of the system matrices or increasing the spectral radius Basabe and Sen [2010]. Another more common method is to decrease the size of elements \(h\). This can also decrease errors according to the size of the grid.
Larson and Bengzon [2013], however the rate at which the error decreases will depend on $p$. This approach also has costs, such as directly increasing the number of degrees of freedom. Often the choice is highly dependent on the mathematical problem being solved numerically.

### 4.1 Higher regularity finite element spaces

Generally most finite element methods for equations up to second-order will use the space $H^1[a, b]$ for a given domain, since further restrictions are often not strictly necessary. However, there may be computational benefits to increasing the regularity. Firstly, it is important to note that the sets associated with the Hilbert-Sobolev spaces are nested inside of each other, with each increase in $r$ further restricting the set:

\[ \cdots \subset H^3[a, b] \subset H^2[a, b] \subset H^1[a, b]. \]  

(4.1)

This is because the criteria for each $r$ automatically fulfill the criteria for $r-1, \ldots, 1$. Furthermore, since there are members of $L^2[a, b]$ that are not in $H^1[a, b]$, by taking successive integrals of these functions it is always possible to find a function in $H^{k-1}[a, b]$ that is not in $H^k[a, b]$ for any $k \in \mathbb{N}$, so these are strict subsets.

Returning to the equations, the example of the wave equation, a time-dependent PDE, is considered. The following was shown for the 1D case in Weber et al. [2022]. A finite element discretisation will generally have the form

\[ M \ddot{u} = -Au, \]  

(4.2)

where $u$ is an array of time-dependent coefficients for the spatial basis functions, and the entries of the mass matrix $M$ and the stiffness matrix $A$ are given by

\[ M_{i,j} = \int_a^b \psi_i(x)\psi_j(x)dx, \quad A_{i,j} = \int_a^b \partial\psi_i(x)\partial\psi_j(x)dx, \]

for spatial basis functions $\psi_i$ where $i$ indexes the basis. Generally, the numerical stability of an explicit time-stepping method will depend on the largest eigenvalue $\lambda$ of $M^{-1}A$, with the largest stable time-step being proportional to $\lambda^{-\frac{1}{2}}$.\(^{10}\)

\(^{10}\)Note that since $M$ is symmetric positive-definite and $A$ is symmetric positive semi-definite any eigenvalues must be real and non-negative, so $\lambda^{-\frac{1}{2}}$ will be a non-negative real value.
The maximum eigenvalue can be written as a maximum of a functional over a function space as follows:

\[
\lambda = \max_{v \in \mathbb{R}^n \setminus \{0\}} \left[ \mu : M^{-1} A v = \mu v \right],
\]

\[
= \max_{v \in \mathbb{R}^n \setminus \{0\}} \left[ \frac{v^T A v}{v^T M v} \right].
\]

\[
= \max_{\eta(x) \in \text{span}\{\psi_i, 1 \leq i \leq n\}} \left[ \frac{\int_a^b |\partial \eta|^2 \, dx}{\int_a^b |\eta|^2 \, dx} \right].
\]

By restricting the space spanned by the basis functions, the maximum value can only decrease. Since this space approximates the Hilbert-Sobolev space \( H^r[a,b] \), this implies that imposing a greater regularity \( r \) will generally lead to a larger value for the maximum stable time-step used in time integration. Numerical experiments found that this restriction could lead to significantly improved maximum time-steps Weber et al. [2022].

References


Part II

Solutions to exercises
Chapter 1

Metric spaces
Ex. 1.1.6

Consider the complex sequence space $l^\infty$ defined as follows:

$$l^\infty = \left\{ x = \{\xi_j\}_{j=1}^{\infty} : \xi_j \in \mathbb{C}, |\xi_j| \leq c_x \forall j \in \mathbb{N} \right\},$$

where $c_x$ can depend on $x$ but not on $j$. Define a function $d : l^\infty \times l^\infty \to \mathbb{R}$ as follows:

$$x = \{\xi_j\}_{j=1}^{\infty}, \quad y = \{\eta_j\}_{j=1}^{\infty}, \quad x, y \in l^\infty \implies d(x, y) := \sup_{j \in \mathbb{N}} |\xi_j - \eta_j|.$$

To show that $(l^\infty, d)$ is a metric space, we consider the four axioms in the definition:

1. From the properties of the norm $|\cdot|$ on $\mathbb{C}$ we have $|\xi_1 - \eta_1| \geq 0$, so it follows that $d(x, y) \geq 0 \forall x, y \in l^\infty$. To prove that $d(x, y)$ is finite, we use the triangle inequality for $\mathbb{C}$:

$$|\xi_j - \eta_j| \leq |\xi_j| + |\eta_j| \leq c_x + c_y \forall j \in \mathbb{N}.$$

Since $|\xi_j - \eta_j|$ is bounded above by a finite limit the supremum must also be less than or equal to this limit. Therefore $d(x, y) \leq c_x + c_y$, so $d(x, y)$ is finite for all $x, y \in l^\infty$.

2. $d(x, y) = 0$ means that the supremum of $|\xi_j - \eta_j|$ is zero. This is true if and only if $|\xi_j - \eta_j| = 0 \forall j \in \mathbb{N}$. This in turn is true if and only if $\xi_j = \eta_j \forall j \in \mathbb{N}$, which is an equivalent expression to $x = y$.

3. $|\xi_j - \eta_j| = |\eta_j - \xi_j|$, so it follows that $d(x, y) = d(y, x) \forall x, y \in l^\infty$.

4. Since the supremum behaves as the maximum value of a function in the limit of a sequence, applying more constraints to the considered sequences can never result in the supremum increasing. Using this the triangle inequality follows from the equivalent inequality for $\mathbb{C}$:

$$d(x, z) + d(z, y) = \sup_{j \in \mathbb{N}} |\xi_j - \zeta_j| + \sup_{k \in \mathbb{N}} |\zeta_k - \eta_k|,$n

$$= \sup_{j,k \in \mathbb{N}} \{ |\xi_j - \zeta_j| + |\zeta_k - \eta_k| \} \geq \sup_{j \in \mathbb{N}} \{ |\xi_j - \zeta_j| + |\zeta_j - \eta_j| \},$$

$$\geq \sup_{j \in \mathbb{N}} |\xi_j - \eta_j| = d(x, y) \quad \forall x, y, z \in l^\infty.$$

Here the inequality holds in the limit because the sequences are bounded.

Therefore $(l^\infty, d)$ is a metric space. □

67
Ex. 1.1.8

We will show that
\[
\tilde{d}(x, y) = \int_{a}^{b} |x(t) - y(t)| \, dt
\]
is a metric on \(C[a,b]\). Clearly, \(\tilde{d}\) is real-valued, non-negative and finite (since the functions and integration interval are bounded). The triangle inequality holds:

\[
\tilde{d}(x, z) + \tilde{d}(z, y) = \int_{a}^{b} |x(t) - z(t)| \, dt + \int_{a}^{b} |z(t) - y(t)| \, dt \\
\geq \int_{a}^{b} |x(t) - y(t)| \, dt = \tilde{d}(x, y)
\]

It is left to show that
\[
\tilde{d}(x, y) = 0 \iff x = y.
\]

Clearly,
\[
x = y \implies \tilde{d}(x, y) = 0.
\]

For the other implication, let \(x, y \in C[a,b]\) and \(x \neq y\). Then there exists \(t_0 \in [a, b]\) such that \(y(t_0) \neq x(t_0)\). Since \(x\) and \(y\) are continuous, there exists a \(\delta > 0\) such that for all \(t \in [a, b]\) such that \(|t_0 - t| < \delta\), \(|x(t) - x(t_0)| < |x(t_0) - y(t_0)|/3 \triangleq \epsilon\) and \(|y(t) - y(t_0)| < \epsilon\). Assume that \(t_0 - \delta > a, t_0 + \delta < b\), otherwise, the interval below can easily be modified. Then, for \(t \in [t_0 - \delta, t_0 + \delta]\),

\[
|x(t) - y(t)| = |(x(t) - x(t_0)) - (y(t) - y(t_0)) + x(t_0) - y(t_0)| > \epsilon,
\]

since \(x(t) - x(t_0)\) and \(y(t) - y(t_0)\) are no bigger than \(\epsilon\) in magnitude and \(|x(t_0) - y(t_0)| = 3\epsilon\).

That gives
\[
\tilde{d}(x, y) \geq \int_{t_0 - \delta}^{t_0 + \delta} |x(t) - y(t)| \, dt \geq 2\delta \epsilon > 0.
\]

Thus,
\[
x \neq y \implies \tilde{d}(x, y) > 0,
\]
and \(\tilde{d}\) is a metric.

Ex. 1.1.8

Consider the set \(C[a,b]\) of real-valued continuous functions on the closed interval \([a,b]\), \(b > a\). For two points \(x, y \in C[a,b]\), define the function \(\tilde{d}(x, y)\) as

\[
\tilde{d}(x, y) = \int_{a}^{b} |x(t) - y(t)| \, dt.
\]
1. By definition $\tilde{d}$ must return a real value when defined. Since $[a, b]$ is a closed interval, let $c_x = \max_{t \in [a, b]} |x(t)|$, $c_y = \max_{t \in [a, b]} |y(t)|$. Then

$$\tilde{d}(x, y) = \int_a^b |x(t) - y(t)| \, dt \geq \int_a^b \min_{s \in [a, b]} |x(s) - y(s)| \, dt \geq \int_a^b 0 \, dt = 0,$$

$$\tilde{d}(x, y) \leq \int_a^b \max_{s \in [a, b]} |x(s) - y(s)| \, dt = (b - a) \max_{s \in [a, b]} |x(s) - y(s)|,$$

$$\leq (b - a) \max_{s \in [a, b]} (|x(s)| + |y(s)|) \leq (b - a) (c_x + c_y).$$

Therefore $\tilde{d}(x, y)$ is real, non-negative and finite for all $x, y \in C[a, b]$.

2. Since $C[a, b]$ is a set of continuous functions, for any $x \in C[a, b]$, $t \in [a, b]$ we have the following property by definition:

$$\forall \epsilon > 0 \exists \delta > 0 \text{ such that } |x(s) - x(t)| < \epsilon \forall s : |s - t| < \delta.$$

For any $x, y \in C[a, b]$ the triangle inequality on real numbers can be used to show that $x - y$ is also continuous:

$$\forall \epsilon : \exists \delta_x > 0 \text{ such that } |x(s) - x(t)| < \frac{1}{2} \epsilon \forall s : |s - t| < \delta_x,$$

$$\exists \delta_y > 0 \text{ such that } |y(t) - y(s)| < \frac{1}{2} \epsilon \forall s : |s - t| < \delta_y,$$

\[
\therefore \text{ Let } \delta = \min \{\delta_x, \delta_y\} > 0, \quad |s - t| < \delta \implies |(x(s) - y(s)) - (x(t) - y(t))| \leq |x(s) - x(t)| + |y(t) - y(s)| < \epsilon.
\]

Suppose there exists some $T \in [a, b]$ such that $x(T) \neq y(T)$. Choose $\epsilon = \frac{1}{2} |x(T) - y(T)|$, and a non-zero $\delta$ satisfying the above continuity property for the function $x - y$. Then

$$\tilde{d}(x, y) = \int_a^b |x(t) - y(t)| \, dt,$$

$$\geq \int_{[\min\{b, T + \delta\}, T - \delta]} |x(t) - y(t)| \, dt \geq \int_{[\max\{a, T - \delta\}, \min\{b, T + \delta\}]} \min_{s \in [\max\{a, T - \delta\}, \min\{b, T + \delta\}]} |x(s) - y(s)| \, dt,$$

$$\geq \min_{s \in [\max\{a, T - \delta\}, \min\{b, T + \delta\}]} |x(s) - y(s)| \cdot (\min \{b - a, b - T + \delta, T - a + \delta, 2\delta\}) \geq \epsilon \cdot \min \{b - a, \delta\} > 0.$$

In the event that $x \neq y$ the function $\tilde{d}(x, y)$ must take a positive non-zero value, therefore $\tilde{d}(x, y) = 0 \iff x = y \forall x, y \in C[a, b]$. This requires that $b > a$, otherwise $\tilde{d}(x, y) = 0$ regardless of whether $x(a) = y(a)$ or not.

3. $\tilde{d}(y, x) = \int_a^b |y(t) - x(t)| \, dt = \int_a^b |x(t) - y(t)| \, dt = \tilde{d}(x, y) \forall x, y \in C[a, b]$. 

69
4. The triangle inequality follows from the inequality for real numbers:
\[
\tilde{d}(x, z) + \tilde{d}(z, y) = \int_a^b |x(t) - z(t)| \, dt + \int_a^b |z(t) - y(t)| \, dt,
\]
\[
= \int_a^b (|x(t) - z(t)| + |z(t) - y(t)|) \, dt,
\]
\[
\geq \int_a^b |x(t) - y(t)| \, dt = \tilde{d}(x, y). \quad \square
\]

**Ex. 1.2.4**

A sequence that converges to 0, but is not in $l^p$ for any $1 \leq p < \infty$ is constructed. Let $A_n$ denote the finite sequence

\[
A_n = \left(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}\right)
\]

and create a sequence by the concatenation

\[
A = (A_1, A_2, A_3, \ldots) \equiv (a_j).
\]

This sequence clearly converges to zero. Consider now a fixed $p \in [1, \infty)$. For all $n > p$, each subsequence of $A$ then contributes to the corresponding sum by

\[
s_n = \sum_{k=1}^{n^n} \left|\frac{1}{n}\right|^p = n^{n-p} > 1.
\]

These terms are evidently bounded from below, so it follows that the sum

\[
\sum_{j=1}^{\infty} |a_j|^p = \sum_{n=1}^{\infty} s_n
\]

diverges, so $A$ is not in any $l^p$.

**Ex. 1.2.4**

We can choose for instance the sequence

\[
(\eta_j) = 1, 1, 1/2, 1/2, 1/2, 1/2, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, 1/3, \ldots,
\]

in which the element $1/n$ appears $2^n$ times. We then get

\[
\sum_{j=1}^{\infty} |\eta_j|^p = \sum_{j=1}^{\infty} \frac{2^j}{j^p} = \sum_{j=1}^{\infty} \frac{2^j}{j^p},
\]

which clearly diverges for $1 \leq p \leq \infty$ as fraction in the sum does not vanish as $j \to \infty$.  

70
Ex. 1.2.11

Suppose that \((X, d)\) is a metric space, and define a function \(\tilde{d} : X \times X \to \mathbb{R}\) by

\[
\tilde{d}(x, y) = \frac{d(x, y)}{1 + d(x, y)} \quad \forall x, y \in X.
\]

1. From the definition of a metric, \(d(x, y) \geq 0 \Rightarrow 1 + d(x, y) \geq 1 \forall x, y \in X\), so \(\tilde{d}(x, y)\) is real, non-negative and finite for all \(x, y \in \mathbb{R}\).

2. \(\tilde{d}(x, y) = 0 \iff d(x, y) = 0 \iff x = y \forall x, y \in X\).

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

4. To prove the triangle inequality we first re-write the expression for \(\tilde{d}(x, y)\) slightly:

\[
\tilde{d}(x, y) = 1 - \frac{1}{1 + d(x, y)},
\]

\[
\tilde{d}(x, z) + \tilde{d}(z, y) = 1 - \frac{1}{1 + d(x, z)} + 1 - \frac{1}{1 + d(z, y)}
= 2 - \frac{2 + d(x, z) + d(z, y)}{1 + d(x, z) + d(z, y) + d(x, z)d(z, y)},
\]

\[
= 1 - \frac{1 - d(x, z)d(y, z)}{1 + d(x, z) + d(z, y) + d(x, z)d(z, y)}
\geq 1 - \frac{1}{1 + d(x, z) + d(z, y) + d(x, z)d(z, y)}
\geq 1 - \frac{1}{1 + d(x, y)} = \tilde{d}(x, y).
\]

Thus \((X, \tilde{d})\) is a metric space. \(\square\)

---

Ex. 1.2.14

We have the two metric spaces \((X_1, d_1)\) and \((X_2, d_2)\) and their cartesian product \(X = X_1 \times X_2\). We will show that the following is a metric on \(X\)

\[
\tilde{d}(x, y) = \sqrt{d_1(x_1, y_1)^2 + d_2(x_2, y_2)^2},
\]

where \(x = (x_1 \in X_1, x_2 \in X_2), y = (y_1 \in X_1, y_2 \in X_2)\). Since \(d_1, d_2\) are metrics, they are finite and real-valued, properties which extend to \(\tilde{d}\) since the square root of sums of finite
and real-valued squares is also finite and real-valued. Further, the square root of a real number is non-negative, and thus (M1) holds.

To show that (M2) holds, we note that if \( \tilde{d}(x,y) = 0 \), then \( d_1 = d_2 = 0 \) must hold, and similarly, if \( d_1 = d_2 = 0 \), then it is easy to see that \( d(x,y) = 0 \). Since, \( d_1, d_2 \) are metric spaces, (M2) holds for them, and \( d_1 = d_2 = 0 \) is true if and only if both \( x_1 = y_1 \) and \( x_2 = y_2 \). Thus \( \tilde{d}(x,y) = 0 \) if and only if \( x = y \), and (M2) holds for \( d \).

For (M3) we have again use (M3) holds for \( d_1, d_2 \), and thus

\[
\tilde{d}(x,y) = d_1(x_1, y_1)^2 + d_2(x_2, y_2)^2 = \sqrt{d_1(\sqrt{x_1, y_1})^2 + d_2(\sqrt{x_2, y_2})^2} = \tilde{d}(y, x),
\]

which shows that (M3) holds for \( \tilde{d} \).

Finally, we let \( z = (z_1 \in X_1, z_2 \in X_2) \in X \), and use that the triangle inequality holds for \( d_1, d_2 \).

First, we simplify the notation by introducing \( a_1 \overset{!}{=} d_1(x_1, z_1), b_1 \overset{!}{=} d_1(z_1, y_1), a_2 \overset{!}{=} d_2(x_2, z_2), b_2 \overset{!}{=} d_2(z_2, y_2) \), then

\[
\tilde{d}(x,y) = \sqrt{d_1(x_1, y_1)^2 + d_2(x_2, y_2)^2} \leq \sqrt{(a_1 + b_1)^2 + (a_2 + b_2)^2}.
\]

Expanding the final expression, and using the Cauchy-Schwarz inequality we continue with

\[
\sqrt{a_1^2 + b_1^2 + a_2^2 + b_2^2 + 2(a_1 b_1 + a_2 b_2)} \leq \sqrt{a_1^2 + a_2^2 + b_1^2 + b_2^2 + 2\sqrt{a_1^2 + a_2^2 + b_1^2 + b_2^2}},
\]

where the right-hand side is equal to

\[
\sqrt{(\sqrt{a_1^2 + a_2^2} + \sqrt{b_1^2 + b_2^2})^2} = \sqrt{a_1^2 + a_2^2 + b_1^2 + b_2^2} = \tilde{d}(x,z) + \tilde{d}(z,y),
\]

which shows that the triangle inequality holds, and \( \tilde{d}(x,y) \) is a metric on \( X \).

---

**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 \( \overline{B}(x_0; r) \).

Consider the space \( \mathbb{N} \) with the metric \( d(x,y) = |x-y| \). Consider the open ball \( B(x_0, 1) \). We have \( B(x_0, 1) = \{x_0\} \). Therefore, \( B(x_0; r) = \{x_0\} \). However, \( \overline{B}(x_0; r) = \{x_0-1, x_0, x_0+1\} \).
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 $\tilde{B}(x_0;r)$.

Consider the metric space $(X,d)$, where $d$ is the discrete metric defined $\forall x, y \in X$ as

$$d(x,y) = \begin{cases} 0, & \text{if } x = y \\ 1, & \text{if } x \neq y. \end{cases}$$

Let $x_0 \in X$ and $r = 1$. Then,

$$B(x_0;1) = \{ x \in X : d(x,x_0) < 1 \} = x_0.$$

An accumulation point $x$ of $B(x_0;1)$ must $\forall \varepsilon > 0$ satisfy $d(x,x_0) < \varepsilon$. If $\varepsilon < 1$ this inequality only holds for $x = x_0$. Thus,

$$\overline{B(x_0;r)} = x_0.$$

Furthermore,

$$\tilde{B}(x_0;1) = \{ x \in X : d(x,x_0) \leq 1 \} = X,$$

so $\overline{B(x_0;1)} \neq \tilde{B}(x_0;1)$.

---

Ex. 1.3.12
Consider the space of bounded functions on an interval $B[a,b]$, $b > a$ with the metric

$$d(x,y) = \sup_{t \in [a,b]} |x(t) - y(t)|.$$

Define the sequence $S = \{s_j\}_{j=1}^{\infty}$ on the interval as follows:

$$s_j = b - 2^{-j} (b - a).$$

This is a monotonically increasing sequence in $[a,b]$ that converges to but never reaches $b$ for any finite $j \in \mathbb{N}$. Using this sequence we define a mapping $T$ from the sequence space $l^\infty$ to a subset $B_S[a,b] \subset B[a,b]$:

$$T \{ \xi_j \}_{j=1}^{\infty} = \left[ y : t \mapsto \begin{cases} \xi_1 & a \leq t < s_1 \\
 \xi_j & s_{j-1} \leq t < s_j, j \in \mathbb{N} \setminus \{1\} \\
 0 & t = b \end{cases} \right].$$

Note that $j = \log_2 \left( \frac{b - a}{b - s_j} \right)$, so for any $t < b$ we have $T \{ \xi_j \}_{j=1}^{\infty} (t) = \xi_{J(t)}$, where $J(t) = 1 + \text{floor} \left[ \log_2 \left( \frac{b - a}{b - t} \right) \right]$. This means $T \{ \xi_j \}_{j=1}^{\infty} (t)$ is defined and bounded for all $t$, so is a member of $B[a,b]$ as required. Furthermore $T \{ \xi_j \}_{j=1}^{\infty} (a) = \xi_1$ and $T \{ \xi_j \}_{j=1}^{\infty} (t)(s_j) = \xi_{j+1}$.
which means there is some \( t \in [a, b] \) for which \( T \{ \xi_j \}_{j=1}^{\infty} (t) = \xi_k \) for every \( k \in \mathbb{N} \). Therefore the mapping \( T \) is injective into \( B[a, b] \).

Consider the set of binary sequences and its image in \( B_S[a, b] \):

\[
X = \left\{ x = \{ \xi_j \}_{j=1}^{\infty} : x \in \ell^\infty, \ \xi_j \in \{0, 1\} \ \forall j \in \mathbb{N} \right\}, \quad Y = \{ y = Tx : x \in X \}.
\]

\( X \) can be viewed as a representation of the power set of natural numbers, and is therefore uncountable. The induced metric \( \bar{d} \) on \( Y \subset B[a, b] \) can be shown to be the discrete metric:

\[
\bar{d}(Tx, Tz) = \sup_{t \in [a, b]} |Tx(t) - Tz(t)| = \max_{j \in \mathbb{N}} |\xi_j - \zeta_j|,
\]

\[
= \begin{cases} 
1 & \exists j \text{ such that } \xi_j \neq \zeta_j, \\
0 & \text{otherwise,} \\
= \begin{cases} 
1 & x \neq z, \\
0 & \text{otherwise,} \\
\end{cases} \\
= \begin{cases} 
1 & Tx \neq Tz, \\
0 & \text{otherwise.} \\
\end{cases}
\end{cases}
\]

Therefore the open balls \( B \left( y, \frac{1}{3} \right) \) in \( B[a, b] \) are non-intersecting for \( y \in Y \).

Since the set \( X \) is uncountable and the mapping \( T \) is injective, it follows that the set \( Y \) is uncountable and therefore the set of non-intersecting open balls \( B \left( y, \frac{1}{3} \right) \) is also uncountable. For a set to be dense in \( B[a, b] \) there must be at least one member inside each open ball, thus it follows that any dense set in \( B[a, b] \) must also be uncountable. Therefore the space \( B[a, b] \) is not separable.

\[\square\]

**Ex. 1.6.14**

Does

\[
d(x, y) = \int_{a}^{b} |x(t) - y(t)| dt
\]

define a metric or pseudometric on \( X \) if \( X \) is (i) the set of all real-valued continuous functions on \([a, b]\), (ii) the set of all real-valued Riemann integrable functions on \([a, b]\)?

(i) The space by \( X \) being the set of all continuous functions on \([a, b]\) and metric \( d \) defined above is a metric space, therefore, a pseudometric space.

This has been proved in Question 1.1.8.

(ii) If \( X \) is the set of all real-valued Riemann integrable functions on \([a, b]\), we show that \((X, d)\) is not a metric space. Consider an example of two distinct Riemann integrable functions \( f_1, f_2 \),

\[
f_1 = \begin{cases} 
0, & \text{if } x \in [a, \frac{a+b}{2}] \\
1, & \text{if } x \in (\frac{a+b}{2}, b]
\end{cases}
\]

\[74\]
\[ f_2 = \begin{cases} 0, & \text{if } x \in \left[ a, \frac{a+b}{2} \right) \\ 1, & \text{if } x \in \left[ \frac{a+b}{2}, b \right] \end{cases} \]

Despite \( f_1 \neq f_2 \), we can see that \( d(f_1, f_2) = 0 \). This makes the axiom (M2) of a metric space not fulfilled.

However, \((X, d)\) is a pseudometric space because it fulfills (M1), (M3), (M4) similarly to Question 1.1.8, but also fulfills (M2*) that \( d(x, x) = 0 \).

\[\text{Ex. 1.6.14}\]

Does
\[ d(x, y) = \int_a^b |x(t) - y(t)| \, dt \]
define a metric or pseudometric on \( X \) if \( X \) is (i) the set of all real-valued continuous functions on \([a, b]\), (ii) the set of all real-valued Riemann integrable functions on \([a, b]\)?

**Solution:** In both cases \( \textbf{M1} \) and \( \textbf{M3} \) obviously hold. \( \textbf{M4} \) also holds in both cases since, for \( z(t) \in X \),
\[ d(x, y) = \int_a^b |x(t) - y(t)| \, dt = \int_a^b |x(t) - z(t) + z(t) - y(t)| \, dt = \int_a^b |x(t) - z(t)| + |z(t) - y(t)| \, dt = d(x, z) + d(z, y). \]
Now consider \( \textbf{M2} \). First, note that \( x(t) = y(t) \) implies \( d(x, y) = 0 \) in both cases. Second, if \( X \) is the set of all real-valued continuous functions we have that
\[ d(x, y) = \int_a^b |x(t) - y(t)| \, dt = 0, \]
implies \( x(t) = y(t) \). See the solution to problem 1.1.8 above. If instead \( X \) is the set of all real-values continuous functions the same is not true. This is illustrated by an example. Let \( a < c < b \) and consider the step functions
\[ x(t) = \begin{cases} 1, & a \leq t \leq c, \\ 0, & c < t \leq b, \end{cases} \quad \text{and} \quad y(t) = \begin{cases} 1, & a \leq t < c, \\ 0, & c \leq t \leq b. \end{cases} \]
Clearly, both \( x(t) \) and \( y(t) \) are Riemann integrable functions, and we have
\[ d(x, y) = \int_a^b |x(t) - y(t)| \, dt = 0. \]
But they are not equal since \( x(c) = 1 \) and \( y(c) = 0 \). To summarize, \( d(x, y) \) defines a metric when \( X \) is the set of all real-valued continuous functions on \([a, b]\), and a pseudometric when \( X \) is the set of all real-valued Riemann integrable functions on \([a, b]\).
Ex. 1.6.14

Does

\[ d(x, y) = \int_a^b |x(t) - y(t)|dt \]

define a metric or pseudometric on \( X \) if \( X \) is (i) the set of all real-valued continuous functions on \([a, b]\), (ii) the set of all real-valued Riemann integrable functions on \([a, b]\)?

(i) Continuous Functions

The given

\[ d(x, y) = \int_a^b |x(t) - y(t)|dt \]

is a metric on the set of all real-valued continuous functions on \([a, b]\) as shown in Question 1.1.8.

(ii) Riemann Integrable Functions

To show whether the defined expression is a metric or a pseudometric on the set of all real-valued Riemann integrable functions on \([a, b]\), we have to check if the properties (M1) to (M4) are fulfilled. We can easily see that (M1), (M3) and (M4) are satisfied as shown in Question 1.1.8 since they do not depend on the continuity of the functions.

Checking (M2) we find that

\[ d(x, y) = \int_a^b |x(t) - y(t)|dt \nRightarrow |x(t) - y(t)| = 0 \forall t \in [a, b] \]

since \( x(t), y(t) \) are not continuous. Hence, \( d(x, y) \) is a pseudometric on the set of all real-valued Riemann integrable functions.
Chapter 2

Normed spaces
Ex. 2.3.10

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

Solution: Suppose that a normed space $X$ has a Schauder basis. Then any element $x \in X$ can be written as

$$x = \sum_{k=1}^{\infty} \alpha_k e_k,$$

where $\alpha_k \in \mathbb{R}$ and $e_k$ is the Schauder basis for $X$. Assume that the basis have been normalized such that $\|e_k\| = 1$ for all $k$. Given any $\epsilon > 0$, there is an $N > 0$ such that

$$\|x - \sum_{k=1}^{N} \alpha_k e_k\| < \frac{\epsilon}{2}.$$

Since $\mathbb{Q}$ is dense in $\mathbb{R}$, there exists a sequence $\beta_k \in \mathbb{Q}$ such that

$$|\alpha_k - \beta_k| < \frac{\epsilon}{2N},$$

for every $k$. Then

$$\|x - \sum_{k=1}^{N} \beta_k e_k\| \leq \|x - \sum_{k=1}^{N} \alpha_k e_k\| + \|\sum_{k=1}^{N} \alpha_k e_k - \sum_{k=1}^{N} \beta_k e_k\|$$

$$< \frac{\epsilon}{2} + \sum_{k=1}^{N} |\alpha_k - \beta_k||e_k|| < \frac{\epsilon}{2} + \frac{\epsilon}{2N} \sum_{k=1}^{N} 1 = \epsilon.$$  \hspace{1cm} (2.0.1)

Equation (2.0.1) shows that every $\epsilon$-neighborhood of $x$ contains an element of the subset

$$M = \left\{ \sum_{k=1}^{N} \beta_k e_k \mid \beta_k \in \mathbb{Q}, N > 0 \right\},$$

hence $\bar{M} = X$ and $M$ is dense in $X$. Additionally, $M$ is countable since the rational numbers are countable. This proves that $X$ is separable.

Remark: If $\alpha_k \in \mathbb{C}$ a similar proof holds with $\beta_k \in \mathbb{Q} + i\mathbb{Q}$.

Ex. 2.3.10

Assume $X$ to be a normed space with a Schauder basis $e_n$. Then there exists a unique sequence of scalars $\alpha_i \in K$ such that

$$\|x - \sum_{i=1}^{k} \alpha_i e_i\| \rightarrow 0 \text{ as } k \rightarrow \infty,$$
for any $x \in X$. We can now rewrite this as

$$||x - \sum_{i=1}^{k} \alpha_i e_i|| = ||x - \sum_{i=1}^{k} \alpha_i ||e_i|| e_i|| = ||x - \sum_{i=1}^{k} \beta_i f_i|| \rightarrow 0 \text{ as } k \rightarrow \infty,$$

where we have defined a new basis in terms of $f_i = \frac{\alpha_i}{||e_i||}$ and $\beta_i = \alpha_i ||e_i||$.

More specifically, for any $x \in X$, and for any $\epsilon > 0$, $\exists$ an $m \in \mathbb{N}$ such that

$$||x - \sum_{i=1}^{k} \beta_i f_i|| < \frac{\epsilon}{2} \quad \forall k > m.$$

Now, we can define the set

$$S = \left\{ \sum_{i=1}^{k} \gamma_i f_i : \gamma_i \in \tilde{K}, k \in \mathbb{N} \right\},$$

where $\tilde{K}$ is a countable dense subset of the scalar field $K$. Because $\alpha_i \in K$ and $\gamma_i \in \tilde{K}$, then for any $\epsilon > 0$, $\exists$ and $\gamma_i$ such that $|\alpha_i - \gamma_i| < \frac{\epsilon}{2k} \quad \forall i = 1, ..., k$.

The triangle inequality then gives

$$||x - \sum_{i=1}^{k} \gamma_i f_i|| \leq ||x - \sum_{i=1}^{k} \alpha_i f_i|| + \sum_{i=1}^{k} ||\alpha_i f_i - \gamma_i f_i||$$

$$\leq \frac{\epsilon}{2} + \sum_{i=1}^{k} |\alpha_i - \gamma_i| ||f_i||$$

$$\leq \frac{\epsilon}{2} + \frac{\epsilon}{2}$$

$$= \epsilon.$$

This shows that there exists an $x' \in S$ within a distance $\epsilon$ of $x$ (as defined by the norm). $S$ is then a countable dense subset of $X$, which means $X$ is separable and completes the proof.

---

**Ex. 2.4.4**

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

For two norms on the vector space $X$ to be equivalent, there exist two real constants $a, b > 0$ such that

$$a||x|| \leq ||x|| \leq b||x||.$$
For these equivalent norms to then induce the same topology, the open sets \((X, \| \cdot \|)\) and \((X, \| \cdot \|_0)\) need to be the same.

To show this, we first consider the mapping \(T_1 : (X, \| \cdot \|) \to (X, \| \cdot \|_0)\) and take an arbitrary \(x_0 \in X\). Then for any \(\varepsilon > 0\), \(\exists \delta = a\varepsilon > 0\) such that \(\|x - x_0\| < \delta\) implies

\[
\frac{1}{a}\|x - x_0\| < \frac{a\varepsilon}{a} = \varepsilon.
\]

Now take \(T_2 : (X, \| \cdot \|_0) \to (X, \| \cdot \|)\) and again an arbitrary \(x_0 \in X\). Then for \(\varepsilon > 0\), \(\exists \delta = b\varepsilon > 0\) such that \(\|x - x_0\|_0 < \delta\) implies that

\[
\frac{1}{b}\|x - x_0\| < \frac{b\varepsilon}{b} = \varepsilon.
\]

Thus, both \(T_1\) and \(T_2\) are continuous and the equivalent norms \(\| \cdot \|\) and \(\| \cdot \|_0\) induce the same topology on \(X\).

---

**Ex. 2.4.4**

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

Consider the equivalent norms \(\| \cdot \|_\alpha\) and \(\| \cdot \|_\beta\), then there exists positive numbers \(a, b\) such that for all \(x \in X\)

\[
a\|x\|_\alpha \leq \|x\|_\beta \leq b\|x\|_\alpha
\]

(2.0.2) holds. Let \((X, \mathcal{T}_\alpha)\) and \((X, \mathcal{T}_\beta)\) be the topological spaces under the norms \(\| \cdot \|_\alpha\) and \(\| \cdot \|_\beta\) respectively. The goal is to show that the open sets in \((X, \mathcal{T}_\alpha)\) and \((X, \mathcal{T}_\beta)\) are the same. Since every open subset is a union of balls, a collection \(\mathcal{T}\) can be defined as the collection of all open balls in \(X\). Therefore, we show that any open ball under the norm \(\| \cdot \|_\alpha\) contains an open ball induced by the norm \(\| \cdot \|_\beta\), and vice versa. Consider the balls centred in \(x_0 \in X\) with any \(r > 0\)

\[
B_\alpha(x_0; r) = \{ x \in X | \|x_0 - x\|_\alpha < r \},
\]

\[
B_\beta(x_0; r) = \{ x \in X | \|x_0 - x\|_\beta < r \}.
\]

From (2.0.2) we have the relationships

\[
\|x\|_\alpha \leq \frac{1}{a}\|x\|_\beta,
\]

\[
\|x\|_\beta \leq b\|x\|_\alpha.
\]

So,

\[
B_\alpha(x_0; r) = \{ x \in X | \|x_0 - x\|_\alpha < r \} \leq \left\{ x \in X | \frac{1}{a}\|x_0 - x\|_\beta < r \right\}
\]

\[
= \{ x \in X | \|x_0 - x\|_\beta < ar \} = B_\beta(x_0; ar),
\]

80
and

\[ B_\beta(x_0; r) = \{ x \in X|\|x_0 - x\|_\beta < r \} \]
\[ \leq \{ x \in X|b\|x_0 - x\|_\alpha < r \} \]
\[ = \left\{ x \in X|\|x_0 - x\|_\alpha \leq \frac{r}{b} \right\} = B_\alpha \left( x_0; \frac{r}{b} \right). \]

By scaling any ball in any of the two norms we can make it contain the other. Hence, any open ball in \( J_\alpha \) will also be found in \( J_\beta \), and vice versa. This proves that the two norms induce the same topology on \( X \).

---

**Ex. 2.5.9**

*If \( X \) is compact and \( M \subset X \) is closed, show that \( M \) is compact.*

**Solution:** Since \( X \) is compact and \( M \subset X \), any sequence \((x_n)\) in \( M \) has a subsequence \( x_{n_k} \) that converges to a point \( x \in X \):

\[ x_{n_k} \to x. \]

Since \( x \) is either in \( M \) or an accumulation point of \( M \), and since \( M \) is closed, we have \( x \in M \). We have shown that every sequence in \( M \) has a convergent subsequence. Thus, \( M \) is compact.

---

**Ex. 2.5.10**

*Let \( X \) and \( Y \) be metric spaces, \( X \) compact, and \( T : X \to Y \) bijective and continuous. Show that \( T \) is a homeomorphism.*

For \( T \) to be a homeomorphism, we need to show that \( T^{-1} : Y \to X \) is continuous, i.e. that \( T^{-1}y_n \to T^{-1}y \).

We take an arbitrary sequence \((y_n)\) with \( y_n \to y \in Y \).

Since \( T \) is bijective, we then have \( Tx_n = y_n \) and \( Tx = y \).

Now since \( X \) is a compact metric space, every sequence \((x_n)\) has a convergent subsequence \((x_{n_k})\). We take the subsequence \( x_{n_k} \to x_0 \in X \) and since \( T \) is continuous, we have that

\[ Tx_{n_k} = y_{n_k} \to Tx_0 \]

where \( y_{n_k} \) is a subsequence of \( y_n \) in \( Y \).

As we originally took \((y_n)\) such that \( y_n \to y \) this implies that

\[ Tx_0 = Tx = y \implies x = x_0. \]

Hence, all subsequences \((x_{n_k}) \to x \) which implies that \( x_n \to x \) which by definition means that \( T^{-1}y_n \to T^{-1}y \) so \( T^{-1} \) is continuous and, thus, \( T \) is a homeomorphism.
Ex. 2.5.10

We will show that if $X, Y$ are metric spaces, $X$ is compact and $T : X \rightarrow Y$ is bijective and continuous, $T$ is a homeomorphism, i.e. $T^{-1}$ is also continuous.

Notice first that $Y$ is compact, since $Y = T(X)$ (due to bijectivity), $T$ is continuous and $X$ is compact. We will now show that the image of an open map under $T$ is open, which is equivalent with $T^{-1}$ being continuous. Let $A$ be an open set in $X$, so that the complement $A^c$ is closed. $A^c$ is compact since closed subsets of compact spaces are compact, and since $T$ is continuous, $T(A^c)$ is also compact, and hence closed. Finally, taking the complement in $Y$ gives an open set $(T(A^c))^c = T(A)$, since $Y = T(X)$. Thus for any open $A \subseteq X$, $T(A)$ is open, so $T^{-1}$ is continuous and $T$ is a homeomorphism.

Ex. 2.5.10

Let $(X; d_1), (Y, d_2)$ be metric spaces and $X$ compact. Suppose there exists a continuous and bijective mapping $T : X \rightarrow Y$. The inverse mapping $T^{-1}$ must exist and be bijective because $T$ is bijective. To show $T^{-1}$ is continuous and therefore a homeomorphism, we assume the opposite and prove a contradiction.

Using the definition of continuity, if $T^{-1}$ is discontinuous then

$$\exists y \in Y, \epsilon > 0 \text{ such that } \forall \delta > 0 \exists \hat{y} \in Y \text{ such that } d_2 (y, \hat{y}) < \delta, \quad d_1 (T^{-1}y, t^{-1}\hat{y}) \geq \epsilon.$$ 

Since we can choose such a $\hat{y}$ for any $\delta > 0$, we construct a sequence $\{y_n\}_{n=1}^{\infty}$ obeying the following:

$$d_2 (y, y_n) < 2^{-n}, \quad d_1 (T^{-1}y, T^{-1}y_n) \geq \epsilon \quad \forall n \in \mathbb{N}.$$ 

The sequence $\{T^{-1}y_n\}_{n=1}^{\infty}$ lies inside the compact set $X$, therefore by definition it is possible to choose a subsequence $\{x_m\}_{m=1}^{\infty}$ that converges to some limit $x$. Since $T$ is continuous:

$$\forall \sigma > 0 \exists \rho > 0 \text{ such that } d_1 (\hat{x}, x) < \rho \implies d_2 (T\hat{x}, Tx) < \sigma,$$

and $\{x_m\}_{m=1}^{\infty}$ converges to $x$:

$$\forall \rho > 0 \exists M \text{ such that } d_1 (x_m, x) < \rho \quad \forall m > M,$$

it follows that $\{Tx_m\}_{m=1}^{\infty}$ converges to $Tx$. However, by definition $\{Tx_m\}_{m=1}^{\infty}$ must be a subsequence of $\{y_n\}_{n=1}^{\infty}$ and must therefore converge to $y$. Since there can be only one limit we require that $T^{-1}y = x$. By the definition of $y_n$ we therefore have $d_1 (x_m, x) > \epsilon > 0$ for all $m \in \mathbb{N}$ if $T^{-1}$ is discontinuous, which is a contradiction. 

\footnote{Consider a sequence in a closed $M \subseteq X$. Since $X$ is compact, the sequence has a subsequence which converges in $X$. Since $M$ contains its limit points, the subsequence converges to a point in $M$.}
**Ex. 2.5.10**

Let $X$ and $Y$ be metric spaces, $X$ compact, and $T : X \to Y$ bijective and continuous. Show that $T$ is a homeomorphism (cf. Prob. 5, Sec. 1.6).

Since $T$ is continuous and bijective, we only need to show that $T^{-1}$ is continuous. For this purpose, we will show that $B = (T^{-1})^{-1}(A)$ is a closed subset of $Y$ for any closed subset $A$ of $X$.

Because $T$ is bijective and hence $T^{-1}$ is also bijective, $(T^{-1})^{-1}(A) = T(A)$. Since $X$ is compact, $A$ is also compact since it is closed. Because $A$ is compact and $T$ is continuous, $B = T(A)$ is a compact set. Therefore, $B$ is closed.

**Ex. 2.7.10**

The norms $\|S\|$, $\|T\|$, $\|ST\|$ and $\|TS\|$ will be calculated for the operators $S, T$ on $C[0, 1]$ defined by

$$
y = Sx, \text{ where } y(t) = t \int_{0}^{1} x(\tau)d\tau,
$$

$$
y = Tx, \text{ where } y(t) = tx(t).
$$

It is easy to verify that $S$ and $T$ are bounded linear operators. The formula

$$
\|\Sigma\| = \sup_{x \in \mathcal{D}(\Sigma), \|x\| = 1} \|\Sigma x\|,
$$

which holds for any bounded linear operator $\Sigma$, can then be used to calculate the norms.

If $y = Sx$ then $y(t) = c_1 t$, where $c_1 = \int_{0}^{1} x(\tau)d\tau$. It then follows that the function $x$ with $\|x\| = 1$ which maximizes $\|Sx\|$ is given by $x(t) = 1$ (or $x(t) = -1$), since that maximizes the integral which determines $c$. That gives

$$
\|S\| = \max_{t \in [0, 1]} |t| = 1.
$$

Since $|tx(t)| \leq 1$ when $t \in [0, 1]$ and max $|x(t)| = 1$, it follows that

$$
\|T\| = 1.
$$

The compositions $ST$ and $TS$ are given by

$$
y = STx, \text{ where } y(t) = t \int_{0}^{1} \tau x(\tau)d\tau,
$$

$$
y = TSx, \text{ where } y(t) = t^2 \int_{0}^{1} x(\tau)d\tau.
$$

We note that $S$ and $T$ do not commute.
Since \( \|TS\| \leq \|T\| \|S\| = 1 \) and \( x = 1 \) gives \( \|TSx\| = 1 \), it follows that
\[
\|TS\| = 1.
\]

Finally, \( y = STx \) gives \( y(t) = c_2 t \), where \( c_2 = \int_0^1 \tau x(\tau)d\tau \), i.e. maximization of \( c_2 \) with \( \|x\| = 1 \) maximizes \( \|STx\| \). \( x(t) = 1 \) again accomplishes this, which gives
\[
\|ST\| = 1/2.
\]

**Ex. 2.10.8**

The dual space of \( c_0 \subset l^\infty \) will be shown to be \( l^1 \). First note that \( (e_k) \) with its usual definition is a Schauder basis for \( c_0 \), so any \( x \in c_0 \) can be uniquely expressed as
\[
x = \sum_{k=1}^{\infty} \xi_k e_k.
\]

Let \( f \in c_0' \) be arbitrary. Since \( f \) is linear and bounded
\[
f(x) = \sum_{k=1}^{\infty} \xi_k \gamma_k, \quad \gamma_k = f(e_k).
\]

Now consider \( x_n \) defined by
\[
\xi_{k}^{(n)} = \begin{cases} 
\frac{|\gamma_k|}{\gamma_k} & \text{if } k \leq n \text{ and } \gamma_k \neq 0 \\
0 & \text{otherwise}
\end{cases}
\]

Then
\[
f(x_n) = \sum_{k=1}^{\infty} \xi_{k}^{(n)} \gamma_k = \sum_{k=1}^{n} |\gamma_k|,
\]

which gives
\[
\sum_{k=1}^{n} |\gamma_k| = f(x_n) \leq \|f\| \|x_n\| = \|f\| \sup_{k} |\xi_{k}^{(n)}| = \|f\|.
\]

Letting \( n \rightarrow \infty \) gives
\[
\sum_{k=1}^{\infty} |\gamma_k| \leq \|f\|, \quad (2.0.3)
\]

so \( (\gamma_k) \in l^1 \).

Now let \( b = (\beta_k) \in l^1 \) and define the linear functional
\[
g(x) = \sum_{k=1}^{\infty} \xi_k \beta_k.
\]
for $x \in c_0$. Since

$$|g(x)| \leq \sum_{k=1}^{\infty} |\xi_k \beta_k| \leq \|x\| \sum_{k=1}^{\infty} |\beta_k|,$$

$g(x)$ is bounded.

Finally, since

$$|f(x)| = \left| \sum_{k=1}^{\infty} \xi_k \gamma_k \right| \leq \|x\| \sum_{k=1}^{\infty} |\gamma_k|,$$

the norm satisfies

$$\|f\| \leq \sum_{k=1}^{\infty} |\gamma_k|,$$

which together with (2.0.3) gives

$$\|f\| = \sum_{k=1}^{\infty} |\gamma_k|.$$

We thus have a linear bijective map from $c_0'$ to $l^\infty$ defined by $f \mapsto c = (\gamma_k)$ which is norm preserving, i.e. the spaces are isomorphic.
Chapter 3

Inner product spaces
Ex. 3.3.2

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

Take any two elements of \( M \), \( x = (\xi_j) \), \( y = (\eta_j) \). The inner product norm of \( \mathbb{C}^n \) is

\[
\langle x, y \rangle = \sum_{j=1}^{n} \xi_j \bar{\eta}_j.
\]

The induced norm and metric is

\[
\|x\| = \sqrt{\langle x, x \rangle} = \sqrt{\sum_{j=1}^{n} |\xi_j|^2}
\]

and

\[
d(x, y) = \|x - y\| = \sqrt{\sum_{j=1}^{n} |\xi_j - \eta_j|^2}.
\]

Completeness

We know that in a finite dimensional vector space all norms are equivalent as well as that equivalent norms define the same topology. Hence, our task is to show that \( M \) is complete with respect to the metric \( d \) above.

Let \( (x_m) \) be a Cauchy sequence in \( M \),

\[
x_m = (\xi_{m1}, \xi_{m2}, \ldots, \xi_{mn}).
\]

Since \( (x_m) \in M \), we have that

\[
\sum_{j=1}^{n} \xi_{mj} = 1.
\]

Take any \( \varepsilon > 0 \). Then, there exists an \( N \) such that

\[
d(x_m, x_k) < \varepsilon, \quad \forall m, k > N.
\]

Since \( \mathbb{C}^n \) is complete, \( (\xi_{mi}) \) converges in \( \mathbb{C}^n \). Define

\[
\xi_i := \lim_{m \to \infty} \xi_{mi}, \quad \forall i.
\]

Take \( x = (\xi_1, \xi_2, \ldots, \xi_n) \in \mathbb{C}^n \). Additionally, \( x \in M \) since

\[
\sum_{i=1}^{n} \xi_i = \sum_{i=1}^{n} \lim_{m \to \infty} \xi_{mi} = \lim_{m \to \infty} \sum_{i=1}^{n} \xi_{mi} = 1.
\]

(3.0.1)
From (3.0.1) we can find \( N_i \) such that
\[
|\xi_{mi} - \xi_i| < \frac{\varepsilon}{\sqrt{n}}, \quad \forall m > N_i.
\]
Let \( \hat{N} = \max_i N_i \). Then,
\[
d(x_m, x) = \sqrt{\sum_{i=1}^{n} |\xi_{mi} - \xi_i|^2} < \sqrt{\sum_{i=1}^{n} \frac{\varepsilon^2}{n}} = \varepsilon, \quad \forall m > \hat{N}.
\]
Thus, every Cauchy sequence in \( M \) converges to a point in \( M \). Moreover, since \( M \) is closed, this proves that \( M \) is complete.

**Convexity**

Let \( x = (\xi_1, \xi_2, \ldots, \xi_n) \in M \), \( y = (\eta_1, \eta_2, \ldots, \eta_n) \in M \), and \( \theta \in [0, 1] \). Consider the convex combination of the elements \( x \) and \( y \),
\[
\theta x + (1 - \theta)y = (\theta \xi_1 + (1 - \theta)\eta_1, \theta \xi_2 + (1 - \theta)\eta_2, \ldots, \theta \xi_n + (1 - \theta)\eta_n).
\]

Moreover,
\[
\sum_{j=1}^{n} (\theta \xi_j + (1 - \theta)\eta_j) = \theta \left( \sum_{j=1}^{n} \xi_j \right) + (1 - \theta) \left( \sum_{j=1}^{n} \eta_j \right) = 1,
\]

which means that \( (\theta x + (1 - \theta)y) \in M \), so \( M \) is convex.

**Vector of minimum norm**

Consider Hölder’s inequality with \( p = q = 2 \),
\[
1 = \sum_{i=1}^{n} \xi_i \leq \sum_{i=1}^{n} |\xi_i| \leq \left( \sum_{i=1}^{n} |\xi_i|^2 \right)^{1/2} \left( \sum_{i=1}^{n} |1|^2 \right)^{1/2} = ||x||\sqrt{n}.
\]

So, \( ||x|| \geq 1/\sqrt{n} \), \( n \neq 0 \), where the minimal norm of \( 1/\sqrt{n} \) is obtained using the vector \( y = (1/n, 1/n, \ldots, 1/n) \in M \) since
\[
||y|| = \sqrt{\sum_{i=1}^{n} |\eta_i|^2} = \sqrt{\sum_{i=1}^{n} \frac{1}{n^2}} = \frac{1}{\sqrt{n}}.
\]
Ex. 3.3.2

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

To show that \( M \) is complete, we first consider the vector

\[
1 = (1, 1, \ldots, 1) \in \mathbb{C}^n
\]

which implies that

\[
\sum \eta_j = \langle y, 1 \rangle.
\]

Since \( \mathbb{C}^n \) is a Hilbert space, we take a Cauchy sequence \( y_n \in M \) such that \( y_n \rightarrow y \in \mathbb{C}^n \). Hence,

\[
\langle y_n, 1 \rangle \rightarrow \langle y, 1 \rangle = 1
\]

which shows that \( M \) is complete.

To show that \( M \) is convex, we consider \( x, y \in M \) such that \( x = (\xi_j) \) and \( y = (\eta_j) \) with \( \sum \xi_j = \sum \eta_j = 1 \) and \( 0 \leq t \leq 1 \). We then have

\[
t \left( \sum \xi_j \right) = t,
\]

\[
(1 - t) \left( \sum \eta_j \right) = 1 - t,
\]

\[
\Rightarrow t \left( \sum \xi_j \right) + (1 - t) \left( \sum \eta_j \right) = t + 1 - t = 1
\]

\[
\Rightarrow tx + (1 - t)y = 1 \in M
\]

which shows that \( M \) is convex.

The last step is then to find the minimum norm in \( M \). To do this we take

\[
y = \frac{1}{n} \mathbf{1} + \left( y - \frac{1}{n} \mathbf{1} \right).
\]

This means that

\[
\| y \|^2 = \left\langle \frac{1}{n} \mathbf{1} + \left( y - \frac{1}{n} \mathbf{1} \right), \frac{1}{n} \mathbf{1} + \left( y - \frac{1}{n} \mathbf{1} \right) \right\rangle
\]

\[
= \left\| \frac{1}{n} \mathbf{1} \right\|^2 + \left\| y - \frac{1}{n} \mathbf{1} \right\|^2 + \frac{1}{n} \left[ \left\langle y - \frac{1}{n} \mathbf{1}, \mathbf{1} \right\rangle + \left\langle y - \frac{1}{n} \mathbf{1}, \mathbf{1} \right\rangle \right].
\]

Since we have that both \( y, \frac{1}{n} \mathbf{1} \in M \) this implies that

\[
\langle y, 1 \rangle = \left\langle \frac{1}{n} \mathbf{1}, 1 \right\rangle.
\]

Hence
\[ \|y\|^2 = \left\| \frac{1}{n}, 1 \right\|^2 \geq \left\| \frac{1}{n} \right\|^2 \]

and, thus, \( y = \frac{1}{n}1 \) is the vector which minimises the norm.

---

**Ex. 3.3.4**

(a) We begin by showing that the space \( M \) is complete and convex. By theorem 3.2-4, a closed subspace \( M \) of a Hilbert space \( X \) is also complete. Furthermore, a subspace of a vector space is convex due to its associated algebra, since for any scalar \( \alpha \in [0, 1] \), and any vectors \( x, y \in M \)

\[ \alpha x + (1 - \alpha)y = \alpha x + y - \alpha y \in M, \]

since both scalar multiplication and vector addition map to elements within the set. A Hilbert space is an inner product space and in theorem 3.3-1 there is no requirement that the inner product space \( X \) is not complete. Thus, the assumptions of theorem 3.3-1 hold and therefore also its conclusions.

b) Instead of using the parallelogram equality to ultimately show that

\[ \|y - y_0\|^2 = 2\delta^2 + 2\sqrt{2} - 2\sqrt{2}\|\frac{1}{2}(y + y_0) - x\|^2, \]

where \( \|x - y\| = \delta = \|x - y_0\| \), it is possible to use Appolonius identity, which states that for \( u, v, w \) in an inner product space

\[ \|w - u\|^2 + \|w - v\|^2 = \frac{1}{2}\|u - v\|^2 + 2\|\frac{1}{2}(u + v) - w\|^2. \]

We multiply by two on both sides and rearrange the identity to

\[ \|u - v\|^2 = 2\|w - u\|^2 + 2\|w - v\|^2 - 2\|\frac{1}{2}(u + v) - w\|^2. \]

Now, the elements, \( x, y, y_0 \) are in an inner product space and we may set \( u = y, v = y_0, w = x \), and we immediately get eq. (2).

---

**Ex. 3.3.4**

(a) Show that the conclusion of Theorem 3.3-1 also holds if \( X \) is a Hilbert space and \( M \in X \) is a closed subspace. (b) How could we use Appolonius' identity (Sec. 3.1, Prob. 5) in the proof of Theorem 3.3-1?

(a) Because \( X \) is a Hilbert space and \( M \) is a closed subspace of \( X \), \( M \) is complete. \( X \) is a Hilbert space meaning it is a complete inner product space. Therefore, the assumptions of Theorem 3.3-1 are fulfilled. The theorem holds as a consequence.
(b) The Appolonius’s identity says
\[ \|z - x\|^2 + \|z - y\|^2 = \frac{1}{2} \|x - y\|^2 + 2 \left\| z - \frac{1}{2}(x + y) \right\|^2. \]

In the existence proof, we can replace the equation to prove \((y_n)\) is a Cauchy sequence with
\[ \|x - y_n\|^2 + \|x - y_m\|^2 = \frac{1}{2} \|y_n - y_m\|^2 + 2 \left\| x - \frac{1}{2}(y_n + y_m) \right\|^2. \]
From this, we have
\[ \|y_n - y_m\|^2 \leq 2(\sigma_n^2 + \sigma_m^2) - 2\sigma^2. \]

In the uniqueness proof, we replace the use of the parallelogram equality with
\[ \|x - y\|^2 + \|x - y_0\|^2 = \frac{1}{2} \|y - y_0\|^2 + 2 \left\| x - \frac{1}{2}(y + y_0) \right\|^2. \]
From this, we get a similar conclusion to using the parallelogram equality,
\[ \|y - y_0\|^2 = 2\sigma^2 + 4\sigma^2 - 4 \left\| \frac{1}{2}(y + y_0) - x \right\|^2. \]

**Ex. 3.3.10**

We will show that in a Hilbert space \(H\), if \(M \subseteq H\) and \(M \neq \emptyset\), then for all closed subspaces \(Y \subseteq H\) such that \(M \subseteq Y\) it holds that \(M^\perp \subseteq Y\).

Suppose \(M\) and \(Y\) have the above properties. Since every element that is orthogonal to all elements in \(Y\) is orthogonal to all elements of \(M\), it follows that
\[ Y^\perp \subseteq M^\perp, \]
and, by the same logic
\[ M^\perp \subseteq Y^\perp. \]
But, since \(Y\) is a closed subspace in a Hilbert space, \(Y = Y^\perp\), so we have
\[ M^\perp \subseteq Y. \]
Ex. 3.3.10

Suppose $\emptyset \neq M \subset H$ is a non-empty subset of Hilbert space $H$, and let $Y \subset H$ be a closed subspace of $H$ that contains $M$. We define the orthogonal and double orthogonal complements of $M$ as follows:

$$M^\perp = \{ x \in H \mid \langle x, m \rangle = 0 \ \forall m \in M \},$$

$$M^{\perp \perp} = \{ y \in H \mid \langle x, y \rangle = 0 \ \forall x \in M^\perp \},$$

where $M \subset M^{\perp \perp}$ follows immediately from the definition. We now show that $M^{\perp \perp}$ is a closed subspace of $H$, and that for any closed subspace $Y \subset H$ we have $M^{\perp \perp} \subset Y$.

That $M^{\perp \perp}$ is a subspace of $H$ follows from the bilinearity of the inner product:

$$y_1, y_2 \in M^{\perp \perp} \implies \overline{\alpha_1} \langle x, y_1 \rangle + \overline{\alpha_2} \langle x, y_2 \rangle = 0 \ \forall \alpha_1, \alpha_2 \in \mathbb{C}, \ x \in M^{\perp \perp},$$

$$\implies \langle x, \alpha_1 y_1 + \alpha_2 y_2 \rangle = 0 \ \forall \alpha_1, \alpha_2 \in \mathbb{C}, \ x \in M^{\perp},$$

$$\implies \alpha_1 y_1 + \alpha_2 y_2 \in M^{\perp \perp} \ \forall \alpha_1, \alpha_2 \in \mathbb{C}.$$  

To show $M^{\perp \perp}$ is closed, we note that it is a subspace of Hilbert space $H$, so any Cauchy sequence $\{y_n\}_{n=1}^\infty$ must have a limit $y \in H$. Using the continuity of the inner product:

$$0 = \langle x, y_n \rangle \to \langle x, y \rangle \ \forall x \in M^{\perp},$$

$$\implies \langle x, y \rangle = 0 \ \forall x \in M^{\perp},$$

therefore $y \in M^{\perp \perp}$ and the subspace $M^{\perp \perp}$ is closed.

Suppose now we have a closed subspace $Y$ such that $M \subset Y \subset H$ and consider the orthogonal complement of $Y$:

$$Y^\perp = \{ x \in H \mid \langle x, y \rangle = 0 \ \forall y \in Y \}.$$  

From $M \subset Y$, this means if $x \in Y^\perp$ we automatically get $\langle x, m \rangle = 0 \ \forall m \in M$, so $Y^\perp \subset M^{\perp \perp}$. Similarly the same logic shows that $M^{\perp \perp} \subset Y^\perp$. By lemma 3.3-6 we have $Y^{\perp \perp} = Y$, so $M^{\perp \perp} \subset Y$. Since $Y$ was an arbitrary closed subspace containing $M$, it therefore follows that $M^{\perp \perp}$ is the smallest closed subspace containing $M$. □

Ex. 3.4.6

Let $x^*$ be the remainder of the $x$ with respect to its components in the orthonormal set

$$x - \sum_{k=1}^n \langle x, e_k \rangle e_k = x^*,$$

Then

$$x - y = \sum_{k=1}^n \langle x, e_k \rangle e_k + x^* - \sum_{i=1}^n \beta_i e_i = \sum_{k=1}^n (\langle x, e_k \rangle - \beta_k) e_k + x^*.$$  

92
Then the norm squared becomes
\[ \|x - y\|^2 = \left\| \sum_{k=1}^{n} (\langle x, e_k \rangle - \beta_k) e_k + x^* \right\|^2 = \sum_{k=1}^{n} (\langle x, e_k \rangle - \beta_k)^2 e_k + x^*, \sum_{m=1}^{n} (\langle x, e_m \rangle - \beta_k) e_m + x^* \].

Opening up the final expression using the distributivity of the inner product gives mixed terms, which vanish due to orthogonality, and terms with matching vectors which do not vanish, and we get
\[ \|x - y\|^2 = \sum_{k=1}^{n} |\langle x, e_k \rangle - \beta_k|^2 e_k + |x^*| = \sum_{k=1}^{n} |\langle x, e_k \rangle - \beta_k|^2 + \|x^*\|^2 \geq \|x^*\|^2. \]

Due to non-negativity of the absolute value squared, the equality is achieved for \( \beta_k = \langle x, e_k \rangle \), which is then the smallest value for the norm squared and therefore also for the norm.

**Ex. 3.4.6**

We begin by writing
\[ \|x - y\| = \|x - \sum_{j=1}^{n} \beta_j e_j\| = \| (x - \sum_{j=1}^{n} \langle x, e_j \rangle) + \sum_{j=1}^{n} (\langle x, e_j \rangle - \beta_j) e_j\|, \]
where in the last step we added and subtracted the term \( \sum_{j=1}^{n} \langle x, e_j \rangle\).

We note that the two terms within the norm on the right-hand side are in fact orthogonal as \( (x - \sum_{j=1}^{n} \langle x, e_j \rangle) e_j \perp e_j \) for \( j = 1, \ldots, n \). This allows us to make use of the Pythagorean theorem by squaring both sides to obtain
\[ \|x - y\|^2 = \| (x - \sum_{j=1}^{n} \langle x, e_j \rangle) + \sum_{j=1}^{n} (\langle x, e_j \rangle - \beta_j) e_j\|^2 \]
\[ = \| (x - \sum_{j=1}^{n} \langle x, e_j \rangle)\|^2 + \sum_{j=1}^{n} |\langle x, e_j \rangle - \beta_j|^2 e_j\|^2. \]

It is then clear that \( \|x - y\| \) does indeed depend on \( \beta_1, \ldots, \beta_n \) and, due to the positivity of the norm, the right-hand side is minimized iff \( \beta_j = \langle x, e_j \rangle \) for all \( j \).
Ex. 3.4.8
Suppose that in inner product space $X$ we have an orthonormal sequence \( \{e_n\}_{n=1}^\infty \), and consider some $x \in X$. Applying the Bessel inequality (theorem 3.4-6) gives

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

For some $m \in \mathbb{N}$, let $I_m$ be the set of indices for which

$$i \in I_m \implies |\langle x, e_i \rangle| > \frac{1}{m}. $$

By definition $||x||$ is finite, therefore the inequality implies

$$||x||^2 \geq \sum_{j=1}^\infty |\langle x, e_j \rangle|^2, $$

$$\geq \sum_{j \in I_m} |\langle x, e_j \rangle|^2, $$

$$\geq \frac{1}{m^2} \sum_{j \in I_m} 1.$$ 

This sum is bounded above, therefore the set $I_m$ must have a finite cardinality, which we label $n_m$. Equality only holds in the above inequality if $n_m = 0$ by the definition of $I_m$, which in turn requires that $x = 0$. Therefore the inequality is strict for $x \neq 0$. Multiplying both sides of the inequality by $m^2$ gives

$$n_m < m^2 ||x||^2, \quad x \in X \setminus \{0\}. \quad \Box$$

Ex. 3.5.6
\((e_j)\) is an orthonormal sequence in a Hilbert space $H$. We will show that if $x, y \in H$ and

$$x = \sum_{j=1}^\infty \alpha_j e_j, \quad \sum_{j=1}^\infty \beta_j e_j$$

then

$$\langle x, y \rangle = \sum_{j=1}^\infty \alpha_j \bar{\beta}_j$$

and the sequence is absolutely convergent.
Plugging in the sums for $x$ and $y$ and utilizing that $\langle e_j, e_k \rangle = \delta_{jk}$ gives

$$\langle x, y \rangle = \left( \sum_{j=1}^{\infty} \alpha_j e_j, \sum_{k=1}^{\infty} \beta_k e_k \right) = \sum_{j=1}^{\infty} \alpha_j \left( \sum_{k=1}^{\infty} \beta_k e_k \right) = \sum_{j=1}^{\infty} \alpha_j \langle e_j, e_j \rangle = \sum_{j=1}^{\infty} \alpha_j \beta_j.$$

Since $x, y \in H$, the sums

$$\sum_{j=1}^{\infty} |\alpha_j|^2, \quad \sum_{j=1}^{\infty} |\beta_j|^2 = \sum_{j=1}^{\infty} |\bar{\beta}_j|^2$$

are convergent. The Cauchy-Schwarz inequality can therefore be applied:

$$\sum_{j=1}^{\infty} |\alpha_j \bar{\beta}_j| \leq \left( \sum_{j=1}^{\infty} |\alpha_j|^2 \right)^{1/2} \left( \sum_{j=1}^{\infty} |\bar{\beta}_j|^2 \right)^{1/2} < \infty,$$

so the sequence is absolutely convergent.

**Ex. 3.5.8**

Let $\{e_j\}_{j=1}^{\infty}$ be an orthonormal sequence in Hilbert space $H$, and define subspace $M = \text{span} \{e_j\}$. Since $\bar{M}$ is the smallest closed subset containing $M$ and $H$ is a Hilbert space, by question 3.3-10 it follows that $\bar{M} \subset M^\perp$, since $M^\perp$ is a closed subspace. We write $x \in \bar{M}$ as

$$x = \hat{x} + \sum_{j=1}^{\infty} \langle x, e_j \rangle e_j,$$

$$\Rightarrow \langle \hat{x}, e_j \rangle = \left( x - \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k, e_j \right) = \langle x, e_j \rangle - \sum_{k=1}^{\infty} \langle x, e_k \rangle \langle e_k, e_j \rangle = 0 \quad \forall j \in \mathbb{N},$$

$$\Rightarrow \hat{x} \in M^\perp.$$

However, since $x$ is in the closure of $M$, we can choose a Cauchy sequence in $M$ converging to it. We can also define a Cauchy sequence converging to the sum:

$$x_n = \sum_{j=1}^{m_n} \alpha^n_j e_j, \quad y_n = \sum_{j=1}^{m_n} \langle x, e_j \rangle e_j, \quad m_n \to \infty \text{ as } n \to \infty.$$
be Cauchy and have a limit \( z \in H \). Additionally, \( z_n \in M \) for every \( n \), so \( z \in \bar{M} \subseteq M^\perp \). However, \( z_n = x_n - y_n \to x - y = \hat{x} \), so \( \hat{x} = z \in M^\perp \). Therefore \( \hat{x} \in M^\perp \cap M^\perp \perp = \{0\} \), so for any \( x \in \bar{M} \) we have

\[
x = \sum_{j=1}^{\infty} (x, e_j) e_j.
\]

---

**Ex. 3.6.6**

Suppose that \( H \) is a separable Hilbert space containing a countable dense subset \( M \). We write \( M \) as a list with the following indexing:

\[
M = \{m_1, m_2, \ldots\},
\]

and apply the Gram-Schmidt process to this listing of \( M \) to obtain a countable orthonormal sequence:

\[
S = \{e_1, e_2, e_3, \ldots\},
\]

where orthonormal vector \( e_n \) is obtained after \( s(n) \geq n \) terms of \( M \). Depending on the set \( M \) this sequence might terminate after a finite number of terms while \( M \) is infinite in size, for example \( H = \mathbb{R}, \ M = \mathbb{Q} \) will give \( S = \{1\} \).

By the definition of a dense subset, for any \( x \in H \) we have

\[
\forall \epsilon > 0 \exists m \in M \text{ such that } ||x - n|| < \epsilon.
\]

Suppose for a given \( \epsilon \) such an \( m \) has index \( k(\epsilon) \) in the indexing of \( m \), and the Gram-Schmidt process up to term \( k \) has generated \( n(k) \) terms. Then we can write

\[
m_{k(\epsilon)} \in \text{span} \{e_1, e_2, \ldots, e_{n(k)}\},
\]

\[
\implies m_{k(\epsilon)} = \sum_{j=1}^{n(k)} \alpha_j e_j, \quad \alpha_1, \ldots, \alpha_{n(k)} \in \mathbb{C},
\]

\[
\implies \left\| x - \sum_{j=1}^{n(k)} \alpha_j e_j \right\| < \epsilon.
\]

Since \( \epsilon > 0 \) was arbitrary, this indicates that for any \( x \in H \) a finite series expansion can be made arbitrarily close to \( x \), therefore the orthonormal basis generated by the Gram-Schmidt process on \( M \) produces a total orthonormal basis. \( \square \)
Ex. 3.6.8

Suppose $F$ is an orthonormal sequence in separable Hilbert space $H$. We wish to find a total orthonormal sequence $\tilde{F}$ that contains $F$. If $F$ is total then $\tilde{F} = F$ satisfies this trivially, so we assume it is not total. That is

$$S = \text{span}(F), \quad \tilde{S} \neq H.$$  

This also means that $S^\perp \neq F$, therefore $S^\perp \neq \emptyset$. $H$ is separable, meaning it must have at least one countable dense subset $M$. Since $\emptyset \neq S^\perp \subset H$, it follows that $M$ is also dense in $S^\perp$. For any $m \in M$, let $m = s + y$, where $s \in \tilde{S}$, $y \in S^\perp$. Then for $z \in S^\perp$ we have

$$||z - m||^2 = \langle z - s - y, z - s - y \rangle,$$

$$= ||z - y||^2 - \langle z - y, s \rangle - \langle s, z - y \rangle + ||s||^2,$$

$$= ||z - y||^2 + ||s||^2 \geq ||z - y||^2.$$  

Since $y$ is the projection of $m$ onto $S^\perp$, it follows that $\Pi M$, the projection of $M$ onto $S^\perp$, is a dense subset of $S^\perp$. Using question 3.6-6 we obtain a total orthonormal sequence $G$ in $S^\perp$.

For any $f \in F$ and $g \in G$ we have $||f|| = ||g|| = 1$ by definition, and $\langle f, g \rangle = 0$ since $F \subset S, G \subset S^\perp$. Therefore we define $\tilde{F}$ by

$$F = \{f_1, f_2, \ldots \}, \quad G = \{g_1, g_2, \ldots \},$$

$$\tilde{F} = \{f_1, g_1, f_2, g_2, \ldots \}.$$  

$\tilde{F}$ is total in $H$, since for any $x \in H$, we can write

$$x = s + y, \quad s \in \tilde{S}, y \in S^\perp,$$

$$s = \sum_{j=1}^{\infty} \sigma_j f_j, \quad y = \sum_{j=1}^{\infty} \gamma_j g_j,$$

$$x = \sum_{j=1}^{\infty} (\sigma_j f_j + \gamma_j g_j),$$

which is a series over all the orthonormal elements of $\tilde{F}$.  


Ex. 3.8.10

Show that the inner product $\langle \cdot, \cdot \rangle$ on an inner product space $X$ is a bounded sesquilinear form $h$. What is $\|h\|$ in this case?

It is obvious that the inner product is a sesquilinear form since it is linear in the first argument and conjugate linear in the second argument. We prove that $h$ is bounded. By the Schwarz inequality, we have

$$|h(x, y)| = |\langle x, y \rangle| \leq \|x\| \|y\|,$$
which means \( h \) fulfills the boundedness inequality \( |h(x, y)| \leq c\|x\|\|y\| \) with \( c = 1 \). By definition, the norm of \( h \) reads
\[
\|h\| = \sup_{x, y \neq 0} \frac{|\langle x, y \rangle|}{\|x\|\|y\|} = 1
\]
because of the above inequality. The equality happens since we can choose \( x, y \) such that \( \|x\| = \|y\| = 1 \).

**Ex. 3.8.14**

Due to similarities between the Hermitian form and the inner product, we can use a proof very similar to the one used for proving the Schwarz inequality for the inner product. Let \( z = x - \alpha y \) for some scalar \( \alpha \). Due to positive semi-definiteness and sesquilinearity, we have
\[
0 \leq h(z, z) = h(x - \alpha y, x - \alpha y) = h(x, x) - \alpha h(y, x) + \alpha \overline{\alpha} h(y, y),
\]
and with \( \overline{\alpha} = \frac{h(y, x)}{h(y, y)} \), we get
\[
0 \leq h(x, x) - h(y, x)h(x, y) = h(x, x) - \frac{h(x, y)}{h(y, y)} h(x, y) = h(x, x) - \frac{|h(x, y)|^2}{h(y, y)},
\]
where we simply used the conjugate symmetry of the Hermitian form. Finally, and again due to the positive semi-definiteness, the inequality is unaffected by multiplication by \( h(y, y) \), and we have
\[
|h(x, y)|^2 \leq h(x, x)h(y, y).
\]

**Ex. 3.9.2**

For any \( x, y \in \mathcal{H} \), we can write
\[
\langle x, y \rangle = \langle T^{-1}Tx, y \rangle,
\]
where we use make use of the fact that \( T \) is bijective to deduce that its inverse exists.

We then (twice) use the definition of the Hilbert-adjoint operator \( T^* \) to re-write this as
\[
\langle T^{-1}Tx, y \rangle = \langle x, T^* (T^{-1})^* y \rangle.
\]
Comparing to the previous equation, we deduce that \( T^* (T^{-1})^* = \mathbb{1} \mathcal{H} \). Multiplying with \( (T^*)^{-1} \) from the left then gives the desired relation \( (T^{-1})^* = (T^*)^{-1} \).
Ex. 3.10.10

Let $X$ be an inner product space and $T : X \to X$ an isometric linear operator. If $\dim X < \infty$, show that $T$ is unitary.

Since $T$ is isometric, for $x, y \in X$, we have

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

This induces that

$$\|T\| = \sup_{x \neq 0} \frac{\|Tx\|}{\|x\|} \leq 1.$$  

Therefore, $T$ is bounded. By Theorem 3.9-2, for the bounded linear operator $T$, the unique adjoint operator $T^*$ exists. Moreover, we have that an isometric mapping is injective by the remark in Definition 1.6-1. On a finite dimensional space, if a mapping is injective, it is also surjective, invertible. Therefore, we can apply the bounded inverse theorem on $T$ to induce that $T^{-1}$ exists, is continuous, bounded. $T^{-1}$ is also linear, therefore $(T^{-1})^*$ exists by Theorem 3.9-2.

We have that

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

Therefore, $T^*T = I$. Now we will show that $TT^* = I$. From the invertibility, we have

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

The first equality is the result of Question 3.9.2. It deduces that

$$(T^*)^{-1}T^{-1} = I.$$  

Inverting both sides yields $TT^* = I$. We have $T^*T = TT^* = I$. The statement is proved.

Ex. 3.10.12

Show that $T$ is normal if and only if $T_1$ and $T_2$ in Prob. 4 commute. Illustrate part of the situation by two-rowed normal matrices.

From Problem 4, we define

$$T = T_1 + iT_2 \quad \text{and} \quad T^* = T_1 - iT_2.$$  

(3.0.2)

For $T$ to be normal we need to show that $TT^* = T^*T$. Using (3.0.2), we find that

$$TT^* = (T_1 + iT_2)(T_1 - iT_2) = T_1^2 + T_2^2 + iT_2(-T_1T_2 + T_2T_1) \quad \text{and}$$

$$T^*T = (T_1 - iT_2)(T_1 + iT_2) = T_1^2 + T_2^2 + iT_2(T_1T_2 - T_2T_1).$$

99
So in order to get $TT^* = T^*T$, we need that

$$-T_1T_2 + T_2T_1 = T_1T_2 - T_2T_1$$

$$\implies T_2T_1 = T_1T_2$$

so we need $T_1$ and $T_2$ to commute for $T$ to be normal.

We can see this if we consider the $2 \times 2$ matrices

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad \text{and} \quad A^* = \bar{A}^T = \begin{bmatrix} \bar{a} & \bar{c} \\ \bar{b} & \bar{d} \end{bmatrix}.$$  

These give

$$AA^* = \begin{bmatrix} a\bar{a} + \bar{b}b & a\bar{c} + \bar{b}d \\ c\bar{a} + \bar{d}b & c\bar{c} + \bar{d}d \end{bmatrix}, \quad A^*A = \begin{bmatrix} a\bar{a} + c\bar{c} & b\bar{a} + d\bar{c} \\ a\bar{b} + c\bar{d} & b\bar{b} + d\bar{d} \end{bmatrix}.$$  

Thus, for $A$ to be normal, we need $|b|^2 = |c|^2$ and $a\bar{c} + b\bar{d} = c\bar{a} + d\bar{b}$ which corresponds to $T_1$ and $T_2$ commuting.

---

**Ex. 3.10.13**

If $T_n : H \to H$ ($n = 1, 2, \cdots$) are normal linear operators and $T_n \to T$, show that $T$ is a normal linear operator.

Firstly, we note that $T$ will be a bounded linear operator. Moreover, since $T_n$ is normal $T_nT_n^* = T_n^*T_n$ holds. We want to show that $TT^* = T^*T$, or equivalently that $\|TT^* - T^*T\| = 0$. Applying the triangle inequality to $\|TT^* - T^*T\|$ gives

$$\|TT^* - T^*T\| \leq \|TT^* - T_nT_n^*\| + \|T_nT_n^* - T^*T\|$$

$$= \|(T^*T - T_n^*T_n)^*\| + \|T_n^*T_n - T^*T\|$$

$$= 2\|T_n^*T_n - T^*T\|,$$

where we in the last equality used that the norm of a bounded linear operator and its corresponding Hilbert-adjoint operator are equal. Furthermore, we find that the norm in the last row of (3.0.3) becomes

$$\|T_n^*T_n - T^*T\| \leq \|T_n^*T_n - T_nT_n^*\| + \|T_n^*T_n - T^*T\|$$

$$\leq \|T_n\| \|T_n - T\| + \|T_n^* - T^*\| \|T\|$$

$$= \|T_n\| \|T_n - T\| + \|T_n^* - T^*\| \|T\| \to 0$$

as $n \to \infty$ since $T_n \to T$. Thus, it follows that

$$\|TT^* - T^*T\| \leq 2\|T_n^*T_n - T^*T\| \to 0$$

as $n \to \infty$, which implies that $TT^* = T^*T$, i.e. that $T$ is a normal operator.
Ex. 3.10.14

Let $S : H_S^1 \to H_S^2$ and $T : H_T^1 \to H_T^2$. The respective Hilbert adjoint operators have swapped range and domain by definition. Then by normality of $S$, $S^*S = SS^*$, the domains of the rightern-most operators in the factor must match, and we have that $H_S^1 = H_S^2$, and similarly for $T$: $H_T^1 = H_T^2$. The property that $ST^* = T^*S$, and $TS^* = S^*T$, in a similar manner implies that $H_T^2 = H_T^1$, and $H_S^2 = H_S^1$, respectively. As such, we have that $H_S^1 = H_T^1 = H_S^2 = H_T^2$, and we can apply every aspect of theorem 3.9-4.

By theorem 3.9-4 (b), and the distributive properties of linear operators we get for all feasible elements $v, w$

$$
\langle (S + T)(S + T)^*v, w \rangle = \langle (S + T)(S^* + T^*)v, w \rangle = \langle SS^* + ST^* + TS^* + TT^*v, w \rangle,
$$

Then, by using normality of $S, T$ and that $ST^* = T^*S$, and $TS^* = S^*T$, we can 'flip' all the operator products as

$$
\langle SS^* + ST^* + TS^* + TT^*v, w \rangle = \langle S^*S + T^*S + S^*T + T^*Tv, w \rangle,
$$

which we may then collect, by linearity, as

$$
\langle S^*S + T^*S + S^*S + T^*Tv, w \rangle = \langle (S^* + T^*)(S + T)v, w \rangle = \langle (S + T)^*(S + T)v, w \rangle,
$$

where theorem 3.9-4 (b) was used again for the final expression. This gives that

$$
\langle ((S + T)^*(S + T) - (S + T)(S + T)^*)v, w \rangle = 0,
$$

and by lemma 3.9-3, the two operator products must be equal:

$$(S + T)^*(S + T) = (S + T)(S + T)^*,$$

which in turn means that the operator sum $S + T$ is normal. Similarly, for the product $ST$, we can use theorem 3.9-4 (g), and the normality directly and 'move' the adjoint operators as

$$
\langle ST(ST)^*v, w \rangle = \langle STT^*S^*v, w \rangle = \langle ST^*TS^*v, w \rangle = \langle T^*SS^*Tv, w \rangle = \langle T^*S^*STv, w \rangle = \langle (ST)^*STv, w \rangle.
$$

By similar arguments as used for the operator sum, lemma 3.9-3 ensures that

$$(ST)^*ST = ST(ST)^*,$$

and thus the operator product $ST$ is also normal.