Tan Bui-Thanh

Adjoints and their Roles in Computational, Sciences, Engineering, and Mathematics

January 16, 2026

Springer

Use the template dedic.tex *together with the Springer document class SVMono for monograph-type books or SVMult for contributed volumes to style a quotation or a dedication at the very beginning of your book in the Springer layout*

Foreword

Use the template *foreword.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your foreword in the Springer layout.

The foreword covers introductory remarks preceding the text of a book that are written by a *person other than the author or editor* of the book. If applicable, the foreword precedes the preface which is written by the author or editor of the book.

Place, month year *Firstname Surname*

Preface

This book has grown out of my research for the past 20 years and teaching for the past 10 years. Though adjoint has been pervasive in vast literature across mathematics, engineering, and sciences disciplines, with possible different names, there is no survey/review articles or books that systematically unify a sufficient broad subset of these developments and applications of adjoint on the same mathematical footing starting from the basic definitions. It is perhaps impossible to accomplish such a task in a coherent manner. Over the years, many of my colleagues and students have asked the same question: "Where can we find these derivations and constructions of adjoint and its applications?". My answer has been: "To the best of my knowledge, there is no article or book that systematically covers adjoint and its wide range of applications, unfortunately". In Fall 2022 on a 6am bus from Oak Hill to UT Austin, Dr. Jon Wittmer, then a former PhD student of mine and now (2023) at Meta, asked "what are other usefulness of adjoint?". I listed several fields and applications that I have encountered to Jon. I then talked to myself: "It is time to write about adjoint and its role in many disciplines of computational engineering, sciences, and mathematics". My goal is to provide a sufficient comprehensive material to answer these questions, or at least a place to find answers to similar questions. As we shall see, only a small subset of the topics in this book is often covered in existing textbooks, and furthermore the beautiful connection and interaction between them through adjoint are not typically exposed.

This book expectantly provides a rigorous and unified perspective on the usefulness of adjoint in variety of topics. As a result, this book could give broader views and better insights into the constructions and applications of adjoint beyond a single field. By establishing general results and then developing materials specific to each application, we bring forth the details on how abstract concepts/definitions can be translated into particular applications and the connections among them. The interdisciplinary nature of the book is thus useful for advanced undergraduate students, graduate students, researchers, engineers, scientists, and mathematicians who would like to study

the positive impact of adjoint in their corresponding fields. More importantly, with constructive expositions I really hope that the readers can further apply existing adjoint methods and/or develop new adjoint methods to new applications/disciplines.

I started learning and deriving adjoint methods during my PhD research in which I needed to solve optimization constrained by Ordinary Differential Equations (ODEs) or Partial Differential Equations (PDEs). The goal was to greedily find a reduced basis to construct a "best" reduced-order model for ODEs/PDEs. Since I chose Newton-type optimization approach, not only did I need adjoint for computing the gradient efficiently, but also I had to compute the Hessian-vector product exactly and efficiently for the conjugate gradient iterations. I was taught the mechanics to derive the gradient using adjoint method, and the Hessian-Vector product from the Newton iterations. I later learned the mathematics and figured out a systematic way to construct adjoint methods for both gradient and Hessian-vector product. These materials, among many others discussed in following, are the personal comprehension and derivations that I would like to share them in this book.

The first time one typically sees adjoint operator is in finite dimension (such as transpose of a matrix) via inner products. The existence and the reason the adjoint defined that way are often omitted. These can be found later in advanced textbooks/classes (such as functional analysis) as the Riesz representation theorem is typically deployed to prove the existence of the adjoint for densely defined operators. However, the Riesz representation theorem in finite dimensional Hilbert space is straightforward to prove (one-line proof as we shall see), and thus there is no need to wait for later advanced courses (or never for non-math majors). Another topic in which the Riesz representation theorem plays a key role is optimization. Here, using the Riesz representation of the Fréchet derivative as gradient allows me to show that it is not harder to develop a single optimization theory (instead of two separate theories as usually done in optimization textbooks/classes) that is applicable for both finite and infinite dimensions (calculus of variation). Unlike many textbooks, this approach then allows me to provide a single Lagrangian multiplier theorem for optimization problems with equality constraints in both finite and infinite dimensions. This is then obvious to see that the adjoint (operator, equation, state) naturally arises in the Lagrangian approach as a part of the optimality condition, though often we do not distinguish the Lagrangian approach and adjoint approach.

A reduced gradient approach then falls out from the adjoint approach when applying to an abstract separable (to be defined) optimization problem. I then work out the rigorous details how the abstract reduced gradient approach looks like for abstract finite dimensional problems. In particular, I will show that the backpropagation approach for computing the gradient in training deep neural networks (DNNs) is nothing more than the adjoint approach for computing the reduced gradient. The adjoint approach, however, reveals the precise role of the adjoint solutions—also known as the Lagrangian

multipliers—of the adjoint equations steming from the first order optimality condition using the reduced space approach. Another important view point that we will exploit here is that the DNN training problem, from the adjoint point of view, is a constrained optimization problem with the forward pass as the forward equations. These systematic materials are not in any current textbooks.

Another critical result for adjoint operator is the closed range theorem (CRT) which exposes the relationship between a linear operator and its adjoint via their null and range spaces. The theorem essentially tells us that we cannot understand a linear operator completely unless we thorouhgly know its adjoint and vice versa. Unfortunately, the closed range theorem is often introduced in graduate math class (such as functional analysis). This is unnecessary when restricted to Hilbert spaces as we only need to introduce the elementary concepts of orthogonal complement and the closure of a set to prove the CRT. This little investment is worthwhile as we will need the CRT at many places, where adjoint plays a key role, including the solvability of linear operator equation, a simple proof of the Lagrangian multiplier theorem, orthogonal projection, least squares problems, the fundamental theorem of linear algebra, the Picard theorem for inverting a compact linear operator, and the well-posedness of linear operator equation. Our unusual exposition not only highlights the vital role of adjoint in these topics, but also brings out the connections among them through adjoint.

The singular value decomposition is perhaps one of the most useful decomposition in modern scientific computing. One of the highlights of this book is the systematic construction of a single singular value decomposition (SVD)—valid in both finite and finite dimensional spaces—and its application in the closed range theorem, the rank-nullity theorem, the fundamental theorem of linear algebra, and the Picard theorem for inverting compact linear operators. We will also show how the SVD can be used as a dimensionality reduction for data processing and model-order reduction. Unlike standard textbooks in which the SVD is typically introduced and constructed in finite dimensional Euclidean spaces, and in which the role of adjoint is bypassed, we expose the essential role of adjoint in a constructive derivation of SVD. This is at the expense of introducing compact operators and the Hilbert-Schmidt theorem on the spectral decomposition of self-adjoint compact operators. To make the material more accessible, we begin with an elementary proof of the spectral theorem in finite dimension (i.e. eigenvalue problem for self-adjoint operators), which is then used to provide a straightforward proof of SVD for abstract linear operator in finite dimensions. The beauty here is that the usual SVD for matrices is then readily available through the matrix representation of linear operators. Another appealing feature of this approach is that, after discussing the Hilbert-Schmidt theorem, the proof for the SVD for abstract compact operators in any dimension is essentially the same as the finite dimensional counterpart. Together with the CRT, the proof of Picard theorem for why inverting a compact operator could be ill-posed is simple,

again heavily replying on the adjoint. This then naturally lends itself to the concept of ill-posed problem by Hadamard, and the Tikhonov regularization technique for making ill-posed problems well-posed. As a by-product, we discuss the relationship between ill-posedness and ill-conditioning.

The ill-posedness of inverting compact linear operators immediately asks for conditions under which inverting a linear operator is well-posed. To answer this question for abstract linear operators between Hilbert spaces, we start with a theorem on the equivalence between the boundedness below of an operator and its injectivity together with its closed range. This relies on the CRT and the open mapping theorem. It is then simple for us to show the Banach-Nečas-Babuška (BNB) theorem which exposes the relationship between the injective, surjective, and closed range properties of a linear operator with those of its adjoint. In particular, the BNB theorem shows the bijectivity of a linear operator is equivalent to that of its adjoint, that is, understanding a linear operator implies understanding its adjoint and vice versa. We then derive the implication of BNB theorem for the well-posedness of linear operator equation in finite dimensions. Our main application of interest is then the well-posedness of linear PDEs (for which we take an opportunity to present a view of Green identities from adjoint perspective. This will be the basic for distributional derivatives that we discuss later). To that end, we provide various examples including transport equation, elliptic equations, and a large class of PDEs of Friedrichs' type (embracing elliptic, parabolic, hyperbolic, and mixed type PDEs). We will see that the Lax-Milgram theorem for the well-posedness of linear coercive operator equations is an easy consequence of the BNB theorem. With this exposition, we see the role of BNB not only in the well-posedness PDEs (which is traditionally presented) but also in the well-posedness of finite dimensional linear operator equations in a unified fashion. I also take the opportunity here to rigorously derive the reduced space approach with proper functional settings for constrained optimization problems governed by elliptic and hyperbolic PDEs—a material that one typically cannot find in textbooks. I also provide a rigorous exposition on the key role the adjoint plays in neural ordinary differential equations (Neural ODEs).

One of an important applications for which the interplay of self-adjointness, the Lax-Milgram theorem, compact operators, and the Hilbert-Schmidt theorem is the backbone is a rigorous theory of Sturm-Liouville problem and generalized Fourier series. This approach is not popular as it also requires the concept of closed operators, which is typically considered as an advanced topics in studying partial differential operators. However, it fits well in this book as the backbone materials are already introduced in the previous chapters. I thus take the opportunity to present a self-contained rigorous theory of Sturm-Liouville problem and generalized Fourier series.

The last part of the book devotes to various topics in computational applied mathematics that relies on adjoint to provide insights or to provide effi-

cient computations. As Mike Giles put it[1]: "the subject of adjoints is huge", the selection of topics in this section is again necessary from my own personal interest—limited to a subset of topics that I have been working on or are exposed to. The first topic is on the role of adjoint in the necessary and sufficient conditions for the stability of system of linear ODEs (and hence system of nonlinear ODEs via linearization). The second topic is on the application of adjoint in error correction. The third topic is on the *a posterior* error estimation via adjoint. The fourth topic is the Gauss-Newton Hessian computation using adjoint for both unconstrained and constrained optimization. The fifth topic is the use of adjoint in computing the exact Hessian-vector product for constrained optimization with equality constraints. The last topic is on the role of adjoint in the construction of the alternative direction method of multipliers (ADMM).

Common statements such as "Let us introduce ..." or "Let us define ..." without a constructive motivation always bug me. My concern is that if I have a new problem or new setting, how could I come up with such a statement? With this in mind, I try to avoid this as much as possible by providing constructive arguments that lead to a new concept or approach or definition.

Austin Texas, *Tan Bui-Thanh*
August 2023

[1] In a private communication.

Acknowledgements

Use the template *acknow.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) if you prefer to set your acknowledgement section as a separate chapter instead of including it as last part of your preface.

Contents

Acronyms

Use the template *acronym.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your list(s) of abbreviations or symbols in the Springer layout.

Lists of abbreviations, symbols and the like are easily formatted with the help of the Springer-enhanced `description` environment.

Part I
Introduction

Use the template *part.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page in the Springer layout.

Chapter 1
A brief history of adjoint

Abstract

Chapter 2
Structure of the book

Abstract

Chapter 3
Who and what classes could this book be for?

Abstract

Chapter 4
Notations and conventions

Abstract

In this book, boldface lowercase letters such as $\boldsymbol{u}$ are reserved for vector-valued functions in $\mathbb{R}^n$, for some integer n. Boldface uppercase letters such as $\boldsymbol{A}$ denote matrices, while script uppercase letters such as $\mathscr{A}$ denote operators and superscript ^{T} denotes the transpose of a matrix or a vector. Bold blackboard upper cases, i.e. $\mathbb{X}$ and $\mathbb{Y}$, are used for spaces and sets. For example, $\mathbb{H}^n(\Omega) := \left\{u \in \mathbb{L}^2(\Omega) : \text{ weak derivative up to order } n \text{ residing in } \mathbb{L}^2(\Omega)\right\}$ are standard $\mathbb{L}^2$-based Sobolev's spaces [3, 102]. Lowercase letters are used for scalar-valued functions. *We also use lowercase letters for results that are valid for both finite and infinite dimensional settings.* Boldface uppercase letters are used as bases for vector spaces. We use $\mathbb{F}$ to denote either the set of real ($\mathbb{R}$) or complex ($\mathbb{C}$) numbers and $\Re$ to denote the operation of taking the real part of a complex number. All spaces are either complex or real. As a result, unless otherwise stated, results are valid for both real and complex settings. Unless otherwise explicitly specified, all spaces are Hilbert spaces endowed with appropriate inner products and the corresponding induced norms. For example, Hilbert space $\mathbb{X}$ is endowed with the inner product $(u,v)_{\mathbb{X}}$ for any $u, v \in \mathbb{X}$, and the induced norm $\|u\|_{\mathbb{X}} = \sqrt{(u,u)_{\mathbb{X}}}$. To be call an inner product, the map $(\cdot,\cdot)_{\mathbb{X}} : \mathbb{X} \times \mathbb{X} \to \mathbb{F}$ must satisfy the following three conditions:

1. $(\cdot,\cdot)_{\mathbb{X}}$ is linear w.r.t the second argument, i.e.,

$$(u, \alpha v + \beta w)_{\mathbb{X}} = \alpha (u,v)_{\mathbb{X}} + \beta (u,w)_{\mathbb{X}}$$

2. $(\cdot,\cdot)_{\mathbb{X}}$ must be symmetric, i.e.,

$$(u,v)_{\mathbb{X}} = \overline{(v,u)_{\mathbb{X}}},$$

   where the overline denotes the complex conjugate.
3. $(\cdot,\cdot)_{\mathbb{X}}$ is positive definite, i.e.,

a. $(u,u)_{\mathbb{X}} \geq 0$ for all $u \in \mathbb{X}$, and
b. $(u,u)_{\mathbb{X}} = 0 \Leftrightarrow u = \theta$, where $\theta \in \mathbb{X}$ is the zero vector/function.

We shall frequently identify the dual of any Hilbert space with itself (we will revisit this after Definition 5.5). We define $\mathcal{L}(\mathbb{X},\mathbb{Y})$ as the space of all linear operators from $\mathbb{X}$ to $\mathbb{Y}$, $\mathcal{B}(\mathbb{X},\mathbb{Y})$ as the space of all bounded linear operators from $\mathbb{X}$ to $\mathbb{Y}$, and $\mathcal{C}(\mathbb{X},\mathbb{Y})$ as the space of all continuous mapping from $\mathbb{X}$ into $\mathbb{Y}$. By $\mathcal{C}^n(\mathbb{X})$ and $\mathcal{C}_0^\infty(\mathbb{X})$ we mean the space of n-times continuously differentiable function on $\mathbb{X}$ and the space of test functions (infinitely differentiable functions with compact support in $\mathbb{X}$). Superscript * denotes either the topological dual spaces or adjoint operator or the conjugate transpose of a matrix (or a vector). Superscript $^\perp$ stands for the orthogonal complement, and by ":=", we mean "is defined as".

Let $\boldsymbol{E} = \{e^1,\ldots,e^n\}$ and $\boldsymbol{G} = \{g^1,\ldots,g^m\}$ be orthonormal[1] bases for $\mathbb{X}$ and $\mathbb{Y}$, respectively. For any $u \in \mathbb{X}$, we denote by $\boldsymbol{u}^{\boldsymbol{E}}$ the unique vector of coordinates of u in $\boldsymbol{E}$, and it is easy to see that $\left(\boldsymbol{u}^{\boldsymbol{E}},\boldsymbol{v}^{\boldsymbol{E}}\right)_{\mathbb{F}^n} = (u,v)_{\mathbb{X}}$. The matrix representation of $\mathscr{A}$ with respect to the bases $\boldsymbol{E}$ and $\boldsymbol{G}$ is denoted as $\boldsymbol{A}^{\boldsymbol{EG}}$. When there is no ambiguity on the bases that we refer to, we simply ignore the superscripts for both coordinate vector and matrix representation. We shall denote the ith element of a vector $\boldsymbol{u}$ as $\boldsymbol{u}(i)$ and the element at the ith row and jth column of a matrix $\boldsymbol{A}$ as $\boldsymbol{A}(i,j)$. We also use $\boldsymbol{u}_i$ to denote $\boldsymbol{u}(i)$ and this will be clear from the context. We will use square brackets to express matrices and vectors with a finite number of components. Unless otherwise stated, vectors with finite number of components are column vectors.

- $\mathbb{N}$ is the set of natural numbers
- $\mathbb{R}$ and $\mathbb{C}$ are the set of real and complex numbers, respectively. We use the generic field notation $\mathbb{F}$ for results/discussions that are valid for both $\mathbb{R}$ and $\mathbb{C}$.
- $\mathbb{L}^p(\Omega)$ with $\Omega \subseteq \mathbb{R}^n$ is the standard $\mathbb{L}^p$ spaces with respect to the Lebesgue measure on Ω.
- $\mathbb{C}(\Omega)$ is the space of continuous functions on Ω with the topology generated by the uniform norm $\|\cdot\|_\infty$
- $\mathbb{H}^m(\Omega)$, with $m \in \mathbb{N}$, denotes $\mathbb{L}^2$-based Sobolev spaces of order m.
- $\mathscr{H}$ denotes a reproducing kernel Hilbert space (RKHS). For historical reasons, we use X as a set over which an RKHS is defined.
- K denotes kernels and also compact sets in different chapters.
- $\mathbb{U},\mathbb{V},\mathbb{X},\mathbb{Y}$ typically denote (Hilbert) vector spaces, and $\mathcal{S}$ denotes a subset of some vector space.

[1] Orthonormality is simply for convenience, but not essential.

Part II
Adjoint operators in finite dimensional Hilbert spaces

This part of the book is organized as follows. We begin with the celebrated Riesz representation theorem and the closed range theorem, upon which we shall develop several applications of adjoint. The first application is in Chapter 6 on the solvability of linear operator equations before solving them. The role of adjoint in the study of eigenvalue problems is given in Chapter 7. In Chapter 11, we employ the classical projection theorem together with the closed range theorem to find the necessary and sufficient condition for the optimality of an abstract linear least squares problem. The singular value decomposition (SVD) is the main subject of Chapter 8. The SVD decomposition is then deployed to provide trivial proofs for the closed range theorem, rank-nullity theorem, and the fundamental theorem of linear algebra for abstract linear operators. Optimization with equality constraints is the main topic of Chapter 9 in which we expose at length the role of adjoint in optimization theory that is valid for both finite and infinite dimensions. We then show that a reduced spaced approach using adjoint reduces to the backpropagation of deep neural networks in Chapter 10.

Chapter 5
Preliminaries

Abstract This chapter presents definitions and results that are useful for the latter chapters. The prerequisites for this chapter are:

- Linear operator/map.
- Continuous linear functionals (the equivalence between the boundedness and continuity of a linear functional)
- Vector spaces and linear algebra (coordinate and matrix representation of vectors and linear operator in certain bases)
- Basics on Hilbert spaces (inner product, orthogonality)
- Basics topology (open and closed sets, the closure of a set)

In this part, unless otherwise stated, we assume that $\mathbb{U}$ and $\mathbb{V}$ are finite dimensional vector spaces, i.e. $\dim \mathbb{U} = n < \infty$ and $\dim \mathbb{V} = m < \infty$, where dim denotes the dimension. Recall that if $\mathscr{A} \in \mathcal{L}(\mathbb{U}, \mathbb{V})$ and $\dim \mathbb{U} < \infty$, then $\mathscr{A} \in \mathcal{B}(\mathbb{U}, \mathbb{V})$. add the summary for each chapter and what a reader could learn from each chapter

5.1 Linear mappings and linear functionals

All the results of this section hold for any dimensions. We provide examples for both finite and infinite dimensions. Readers who are interested in only finite-dimensional results can simply view all spaces as finite-dimensional ones.

Definition 5.1 (Linear transformation/map/operator). Consider two inner product vector spaces, $\mathbb{U}$, $(\cdot,\cdot)_{\mathbb{U}}$ and $\mathbb{V}$, $(\cdot,\cdot)_{\mathbb{V}}$. Suppose $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ satisfies the following

$$\mathscr{A}(\alpha u + \beta w) = \alpha \mathscr{A}(u) + \beta \mathscr{A}(w), \tag{5.1}$$

where $\alpha, \beta \in \mathbb{C}$, and $u, w \in \mathbb{U}$. Then, we call[1] $\mathscr{A}$ a linear transformation or map, or an operator from $\mathbb{U}$ to $\mathbb{V}$. *A collection of all linear operators mapping* $\mathbb{U}$ *into* $\mathbb{V}$ *is denoted as* $\mathcal{L}(\mathbb{U}, \mathbb{V})$.

Convention 5.1. *1. For linear operator* $\mathscr{A}$*, we write:* $\mathscr{A}u := \mathscr{A}(u)$.
2. The domain of $\mathscr{A}$ *is defined as*

$$\mathsf{D}(\mathscr{A}) := \{u \in \mathbb{U} : \mathscr{A}(u) \text{ is well-defined}\} \subset \mathbb{U}.$$

Note that we require $\mathsf{D}(\mathscr{A})$ *to be a vector space for the definition of a linear operator* (5.1) *to make sense.*
3. The range of $\mathscr{A}$ *is defined as*

$$\mathsf{R}(\mathscr{A}) := \{\mathscr{A}(u) : u \in \mathsf{D}(\mathscr{A})\} = \{v \in \mathbb{V} : \exists u \in \mathsf{D}(\mathscr{A}) \quad \text{and} \, v = \mathscr{A}(u)\} \subset \mathbb{V}.$$

Note that we shall also use $\mathscr{A}(\mathsf{D}(\mathscr{A}))$ *to denote* $\mathsf{R}(\mathscr{A})$.
4. The kernel of $\mathscr{A}$ *(or the null space of* $\mathscr{A}$*) is defined as*

$$\mathsf{N}(\mathscr{A}) := \{u \in \mathbb{U} : \mathscr{A}(u) = \theta\},$$

where, throughout the paper, θ *denotes either "zero" function or "zero" vector in the appropriate space.*

Remark 5.1. Note that Definition 5.1 and Convention 5.1 are valid for both finite and infinite dimensions. It is an easy exercise to show that $\mathsf{R}(\mathscr{A})$ and $\mathsf{N}(\mathscr{A})$ are vector spaces (see Problem 5.1).

Matrix, integral, and differential operators are linear operators as we demonstrate in the following examples.

Example 5.1. Let us consider $\mathbb{U} = \mathbb{R}^n, \mathbb{V} = \mathbb{R}^m$, and let $\boldsymbol{A} : \mathbb{U} \mapsto \mathbb{V}$ be an $\mathbb{R}^{m \times n}$ matrix. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{U}$, $\alpha, \beta \in \mathbb{F}$, we have

$$\boldsymbol{A}(\alpha \boldsymbol{u} + \beta \boldsymbol{v}) = \alpha \boldsymbol{A}\boldsymbol{u} + \beta \boldsymbol{A}\boldsymbol{v}.$$

Thus, any matrix is a linear operator.

Example 5.2. Consider the space of square integrable functions $\mathbb{L}^2(0,1) := \left\{ f : \int_0^1 |f(t)|^2 \, dt < \infty \right\}$. Let $\mathscr{A} : \mathbb{U} = \mathbb{L}^2(0,1) \to \mathbb{V} = \mathbb{R}$ be defined such that for any $f(t) \in \mathbb{L}^2(0,1)$ and a given function $\omega(t) \in \mathbb{L}^2(0,1)$, we have

$$\mathscr{A} f = \int_0^1 \omega(t) f(t) \, dt.$$

Clearly, for $\alpha, \beta \in \mathbb{F}$:

[1] Though this may not be universal, transformation, map, and operator are used interchangeably in this book for simplicity.

$$\begin{aligned}\mathscr{A}\left(\alpha f(t)+\beta g(t)\right) &= \int_0^1 \omega(t)\left[\alpha f(t)+\beta g(t)\right] dx \\ &= \alpha\int_0^1 \omega(t) f(t)\, dt + \beta\int_0^1 \omega(t) g(t)\, dt \\ &= \alpha\mathscr{A}\left(f(t)\right) + \beta\mathscr{A}\left(g(t)\right),\end{aligned}$$

and thus integrals are linear operators.

Example 5.3. Consider $\mathbb{U}=\mathbb{V}=\mathbb{L}^2(0,1)$ and $\mathscr{A}:\mathsf{D}(\mathscr{A})\subset\mathbb{U}\to\mathbb{V}$ such that

$$\mathscr{A}u=\frac{d^2}{dt^2}u(t).$$

For $\alpha,\beta\in\mathbb{F}$, we have

$$\begin{aligned}\mathscr{A}(\alpha u+\beta v) &= \frac{d^2}{dt^2}\left(\alpha u(t)+\beta v(t)\right) \\ &= \alpha\frac{d^2}{dt^2}u(t)+\beta\frac{d^2}{dt^2}v(t) \\ &= \alpha\mathscr{A}u+\beta\mathscr{A}v.\end{aligned}$$

Thus, differentiation is a linear operator.

Definition 5.2 (Bounded linear operator). A linear operator $\mathscr{A}:\mathbb{U}\to\mathbb{V}$ is called bounded if there exists a constant $\alpha>0$ such that

$$(\mathscr{A}u,v)_{\mathbb{V}}\le\alpha\left\|u\right\|_{\mathbb{U}}\left\|v\right\|_{\mathbb{V}},\quad\forall u\in\mathbb{U}\text{ and }v\in\mathbb{V}.$$

When $\mathscr{A}$ is bounded, we define the operator norm of $\mathscr{A}$ as

$$\left\|\mathscr{A}\right\|:=\sup_{u\in\mathbb{U},v\in\mathbb{V}}\frac{(\mathscr{A}u,v)_{\mathbb{V}}}{\left\|u\right\|_{\mathbb{U}}\left\|v\right\|_{\mathbb{V}}}=\sup_{u\in\mathbb{U}}\frac{\left\|\mathscr{A}u\right\|_{\mathbb{V}}}{\left\|u\right\|_{\mathbb{U}}}.\tag{5.2}$$

We would like to point out that, in the definition of $\|\mathscr{A}\|$, while the first equality is the definition, the second one is due to Cauchy-Schwarz inequality. *A collection of all bounded linear operators mapping from* $\mathbb{U}$ *into* $\mathbb{V}$ *is denoted as* $\mathcal{B}(\mathbb{U},\mathbb{V})$*, and clearly* $\mathcal{B}(\mathbb{U},\mathbb{V})\subseteq\mathcal{L}(\mathbb{U},\mathbb{V})$.

Definition 5.3 (Continuous linear operator). A linear operator $\mathscr{A}:\mathbb{U}\to\mathbb{V}$ is continuous at u_0 if $\forall\varepsilon>0$, $\exists\delta(\varepsilon):\left\|u-u_0\right\|_{\mathbb{U}}<\delta\implies\left\|\mathscr{A}(u)-\mathscr{A}(u_0)\right\|_{\mathbb{V}}<\varepsilon$.

Remark 5.2. Clearly, Definition 5.3 is equivalent to the following statement: $\mathscr{A}$ is continuous at u_0 iff $u_n\to u\in\mathbb{U}\implies\mathscr{A}u_n\to\mathscr{A}u\in\mathbb{V}$.

It turns out that linear mappings with finite-dimensional domain are always continuous.

Proposition 5.1. *$\mathscr{A} \in \mathcal{L}(\mathbb{U}, \mathbb{V})$, and $\dim \mathbb{U} < \infty$. Then, $\mathscr{A} \in \mathcal{B}(\mathbb{U}, \mathbb{V})$.*

Proof. Suppose $\dim \mathbb{U} = n$ and let $\boldsymbol{E} = \{e^1, \ldots, e^n\}$ be an orthonormal basis of $\boldsymbol{E}$. Consider an arbitrary $u \in \mathbb{X}$ and its coordinate vector $\boldsymbol{u} \in \mathbb{R}^n$ with respect to $\boldsymbol{E}$. We have

$$\begin{aligned}\|\mathscr{A}u\|_{\mathbb{Y}} = \left\|\sum_{i=1}^{n} \boldsymbol{u}_i \mathscr{A}e^i\right\| \leq \sum_{i=1}^{n} |\boldsymbol{u}_i| \left\|\mathscr{A}e^i\right\|_{\mathbb{Y}} \leq \sqrt{n} \|\boldsymbol{u}\|_{\mathbb{R}^n} \max_{i=1,\ldots,n} \left\|\mathscr{A}e^i\right\|_{\mathbb{Y}} \\ = \sqrt{n} \|u\|_{\mathbb{U}} \max_{i=1,\ldots,n} \left\|\mathscr{A}e^i\right\|_{\mathbb{Y}},\end{aligned}$$

where we have used the triangle inequality in the first inequality, the Cauchy-Schwarz inequality for $\mathbb{R}^n$ in the second inequality, and $\|u\|_{\mathbb{U}} = \|\boldsymbol{u}\|_{\mathbb{R}^n}$ in the last equality. It follows that

$$\|\mathscr{A}\| \leq \sqrt{n} \max_{i=1,\ldots,n} \left\|\mathscr{A}e^i\right\|_{\mathbb{Y}},$$

and this concludes the proof.

The next result is an important characterization of bounded linear operators.

Lemma 5.1 (Continuous linear operator ≡ bounded linear operator). *Let $\mathscr{A} \in \mathcal{L}(\mathbb{U}, \mathbb{V})$. Then, the following are equivalent:*

L1) *$\mathscr{A}$ is continuous at $\theta \in \mathbb{U}$.*
L2) *$\mathscr{A}$ is continuous everywhere.*
L3) $\exists \alpha > 0 : \|\mathscr{A}u\|_{\mathbb{V}} \leq \alpha \|u\|_{\mathbb{U}}, \quad \forall u \in \mathbb{U}.$

Proof. The fact that **L2)** $\implies$ **L1)** is trivial. We shall first prove **L1)** $\implies$ **L3)**. From Definition 5.3, $\mathscr{A}$ is continuous at $\theta \in \mathbb{U}$ implies:

$$\forall \varepsilon, \quad \exists \delta = \delta(\varepsilon), \quad \|u\|_{\mathbb{U}} \leq \delta \implies \|\mathscr{A}u\|_{\mathbb{V}} \leq \varepsilon, \tag{5.3}$$

by definition of continuity. Now pick $\varepsilon = 1$ then for any $w \neq 0$, we have:

$$\left\| \underbrace{\frac{\delta}{2} \frac{w}{\|w\|_{\mathbb{U}}}}_{u:=} \right\|_{\mathbb{U}} = \frac{\delta}{2} < \delta.$$

Then,

$$\left\| \frac{\delta}{2} \frac{\mathscr{A}w}{\|w\|_{\mathbb{U}}} \right\| = \|\mathscr{A}u\|_{\mathbb{V}} \leq 1$$

owing to (5.3). This implies:

$$\|\mathscr{A}w\|_{\mathbb{V}} \leq \underbrace{2/\delta}_{\alpha:=} \|w\|_{\mathbb{U}}, \quad \forall w \in \mathbb{U}.$$

Next, we show **L3)** $\implies$ **L2)**. Consider the sequence $u_n \to u \in \mathbb{U}$. Then **L3)** implies:

$$\|\mathscr{A}(u-u_n)\|_{\mathbb{V}} \leq \alpha \|u-u_n\|_{\mathbb{U}} \to 0 \implies \mathscr{A}u_n \to \mathscr{A}u,$$

that is, $\mathscr{A}$ is continuous at any u.

When both $\mathbb{U}$ and $\mathbb{V}$ are finite dimensional, a linear map $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ is automatically continuous.

Proposition 5.2. *Let $\mathscr{A} \in \mathcal{L}(\mathbb{U}, \mathbb{V})$, and $\dim \mathbb{U}, \dim \mathbb{V} < \infty$. Then $\mathscr{A}$ is continuous.*

Proof.

Definition 5.4 (Linear functional). When the range of a linear operator $\mathscr{L} : \mathbb{U} \to \mathbb{V}$ is a scalar field $\mathbb{F}$ (either real $\mathbb{R}$ or $\mathbb{C}$), we call $\mathscr{L}$ a linear functional.[2]

Remark 5.3. Clearly, all the above results for linear operators are also valid for linear functionals.

Definition 5.5 ((Topological) Dual spaces). We call $\mathcal{B}(\mathbb{U}, \mathbb{F})$ the dual space of $\mathbb{U}$ and denote it as $\mathbb{U}^*$.

It turns out that $\mathbb{U}$ and $\mathbb{U}^*$ is topologically equivalent in the sense that there is a bijective continuous mapping, known as the Riesz map $\mathcal{R}$, between $\mathbb{U}$ and $\mathbb{U}^*$. The injectivity and unitary of the Riesz map $\mathcal{R}$ are left as an exercise in Problem 5.12. The surjectivity of the Riesz map is the content is the following theorem. $\mathbb{U}$ and $\mathbb{U}^*$ is thus isometrically equivalent.

Theorem 5.1 (Riesz representation theorem). *Let $\mathscr{L} : \mathbb{U} \to \mathbb{F}$ be a bounded linear functional on a Hilbert space $\mathbb{U}$. There exists a unique $\ell \in \mathbb{U}$ such that*

$$\mathscr{L}(v) = (\ell, v)_{\mathbb{U}}, \quad \forall v \in \mathbb{U}. \tag{5.4}$$

Furthermore, the operator norm of $\mathscr{L}$ is given as $\|\mathscr{L}\| := \sup_{v \in \mathbb{U}} \dfrac{|\mathscr{L}(v)|}{\|v\|_{\mathbb{U}}} = \|\ell\|_{\mathbb{U}}$.

Proof. A general proof that works for both finite and infinite dimensional settings is quite standard and can be found in any functional analysis book (see, e.g., [112, 11, 28, 97, 131]). We provide a short and intuitive proof for finite dimensions. We prove the result for $\mathbb{F} = \mathbb{C}$ as the case $\mathbb{F} = \mathbb{R}$ is analogous. Let $\boldsymbol{u} = \{e_1, \ldots, e_n\}$ be an orthonormal basis for $\mathbb{U}$. Let $\boldsymbol{u}$ be the representation of u in $\boldsymbol{u}$. we have

$$u = \sum_{i=1}^{n} \boldsymbol{u}(i)\, e_i \implies \mathscr{L}u = \sum_{i=1}^{n} \boldsymbol{u}(i) \underbrace{\mathscr{L}e_i}_{=:\boldsymbol{\ell}(i)} = (\boldsymbol{\ell}, \boldsymbol{u})_{\mathbb{F}^n} = (\ell, u)_{\mathbb{U}}, \quad \forall u \in \mathbb{U},$$

[2] Thus, a linear functional a special linear operator whose range is a scalar field.

where we have defined ℓ through its coordinate vector $\boldsymbol{\ell}$ in the basis $\boldsymbol{u}$, and thus it is unique. We have

$$\|\mathscr{L}\| := \sup_{v\in\mathbb{U}} \frac{|\mathscr{L}(v)|}{\|v\|_{\mathbb{U}}} = \sup_{v\in\mathbb{U}} \frac{|(\ell, v)_{\mathbb{U}}|}{\|v\|_{\mathbb{U}}} \leq \sup_{v\in\mathbb{U}} \frac{\|v\|_{\mathbb{U}} \|\ell\|_{\mathbb{U}}}{\|v\|_{\mathbb{U}}} = \|\ell\|_{\mathbb{U}},$$

where we have used Cauchy-Schwarz inequality. The equality happens when $v = \ell$.

Remark 5.4. It is a common practice to *not distinguish* $\mathscr{L} \in \mathbb{U}^*$ and its presentation $\ell \in \mathbb{U}$. The reason is that though they are different in nature, as far as the results of their actions are concerned, they are the same by (5.4). *This is the reason one typically identifies* $\mathbb{U}^*$ *with* $\mathbb{U}$ *in practice.* In particular, we write

$$\ell(v) = \langle \ell, v\rangle_{\mathbb{U}^*\times\mathbb{U}} = (\ell, v)_{\mathbb{U}},$$

where $\langle \ell, v\rangle_{\mathbb{U}^*\times\mathbb{U}}$ is known as the duality pairing. Furthermore, let us denote equip $\mathbb{U}^*$ with the following canonical inner product

$$(f, g)_{\mathbb{U}^*} := \overline{(\mathcal{R}^{-1}f, \mathcal{R}^{-1}g)_{\mathbb{U}}}, \quad \forall f, g \in \mathbb{U}^*, \tag{5.5}$$

where $\mathcal{R}^{-1} : \mathbb{U}^* \to \mathbb{U}$ is the inverse of the Riesz map, and, for example, $\mathcal{R}^{-1}f$ is the Riesz representation of f in $\mathbb{U}$. Then, $\mathbb{U}^*$ is a Hilbert space with the induced norm exactly the operator norm defined in (5.4) (see Problem 5.13). . After the definition of adjoint in Definition 5.6, we can see from Problem 5.12 that $\mathcal{R}^{-1} = \mathcal{R}^*$ since $\mathcal{R}$ is a unitary operator.

Example 5.4 (Fréchet derivative in $\mathbb{R}^n$*).* Let $f : \mathbb{R}^n \to \mathbb{R}$ define the map $\mathscr{D}f(\boldsymbol{u}, \cdot) : \mathbb{R}^n \to \mathbb{R}$ associated with f at any $\boldsymbol{u}$ via

$$\mathscr{D}f(\boldsymbol{u}, \boldsymbol{h}) := (\nabla f(\boldsymbol{u}), \boldsymbol{h})_{\mathbb{R}^n}, \quad \forall \boldsymbol{h} \in \mathbb{R}^n.$$

Clearly $\mathscr{D}f(\boldsymbol{u}, \cdot)$ is linear. $\mathscr{D}f(\boldsymbol{u}, \cdot)$ is also bounded since

$$\|\mathscr{D}f(\boldsymbol{u}, \cdot)\| = \sup_{\boldsymbol{h}\in\mathbb{R}^n} \frac{\mathscr{D}f(\boldsymbol{u}, \boldsymbol{h})}{\|\boldsymbol{h}\|_{\mathbb{R}^n}} = \sup_{\boldsymbol{h}\in\mathbb{R}^n} \frac{(\nabla f(\boldsymbol{u}), \boldsymbol{h})_{\mathbb{R}^n}}{\|\boldsymbol{h}\|_{\mathbb{R}^n}} = \|\nabla f(\boldsymbol{u})\|_{\mathbb{R}^n} < \infty.$$

Now, by the Taylor expansion, we have

$$f(\boldsymbol{u} + \boldsymbol{v}) = f(\boldsymbol{h}) + (\nabla f(\boldsymbol{u}), \boldsymbol{h})_{\mathbb{R}^n} + o(\|\boldsymbol{v}\|_{\mathbb{R}^n}) = f(\boldsymbol{h}) + \mathscr{D}f(\boldsymbol{u}, \boldsymbol{h}) + o(\|\boldsymbol{v}\|_{\mathbb{R}^n}),$$

and thus by Definition 9.2, $\mathscr{D}f(\boldsymbol{u}, \boldsymbol{h})$ is the Fréchet derivative of f at $\boldsymbol{u}$. Moreover, by Theorem 5.1, we conclude that $\nabla f(\boldsymbol{u})$ is the Riesz representation of $\mathscr{D}f(\boldsymbol{u}, \cdot)$. A general definition of Fréchet derivative for mappings between arbitrary Hilbert space is referred to section 9.2.

5.2 Adjoint operators

Definition 5.6 (Adjoint operator). Let $\mathscr{A} \in \mathcal{B}(\mathbb{U}, \mathbb{V})$. We say that $\mathscr{A}^* : \mathbb{V} \to \mathbb{U}$ is the adjoint of A iff

$$(\mathscr{A}u, v)_{\mathbb{V}} = (u, \mathscr{A}^* v)_{\mathbb{U}}, \quad \forall u \in \mathbb{U}, v \in \mathbb{V}.$$

Proposition 5.3. *Let $\mathscr{A} \in \mathcal{B}(\mathbb{U}, \mathbb{V})$. Then $\mathscr{A}^*$ exists and is unique. Furthermore, it is linear with $\|\mathscr{A}^*\| = \|\mathscr{A}\|$, where the operator norm is defined as in (5.2), e.g.,*

$$\|\mathscr{A}\| := \sup_{u \in \mathbb{U}} \frac{\|\mathscr{A}u\|_{\mathbb{V}}}{\|u\|_{\mathbb{U}}} = \sup_{\|u\|_{\mathbb{U}}=1} \|\mathscr{A}u\|_{\mathbb{V}}. \tag{5.6}$$

Proof. To see the existence and linearity, we note that owing to the continuity of $\mathscr{A}$ and the inner product, $(\mathscr{A}u, v)_{\mathbb{V}}$ is continuous in u. By the Riesz representation Theorem 5.1, there exists a unique $\ell \in \mathbb{U}$ depending on $\mathscr{A}$ and v such that

$$\overline{(\mathscr{A}u, v)_{\mathbb{V}}} = (v, \mathscr{A}u)_{\mathbb{V}} = (\ell(\mathscr{A}, v), u)_{\mathbb{U}}, \quad \forall u \in \mathbb{U},$$

which implies that $\ell(\mathscr{A}, v)$ is linear in v. Defining $\mathscr{A}^*$ via $\mathscr{A}^* v := \ell(\mathscr{A}, v)$ implies the existence, linearity, and uniqueness of $\mathscr{A}^*$. We next show that $\mathscr{A}^*$ is continuous and $\|\mathscr{A}^*\| = \|\mathscr{A}\|$. We have

$$\|\mathscr{A}^* v\|_{\mathbb{U}}^2 = (\mathscr{A}^* v, \mathscr{A}^* v)_{\mathbb{U}} = (\mathscr{A}\mathscr{A}^* v, v)_{\mathbb{V}} \leq \|\mathscr{A}\| \, \|\mathscr{A}^* v\|_{\mathbb{U}} \, \|v\|_{\mathbb{V}},$$

which shows that $\mathscr{A}^*$ is bounded and $\|\mathscr{A}^*\| \leq \|\mathscr{A}\|$. Since $\mathscr{A}$ is the adjoint of $\mathscr{A}^*$, following a similar arguement we have $\|\mathscr{A}\| \leq \|\mathscr{A}^*\|$, and this concludes the proof.

Remark 5.5. From the proof of Theorem 5.1 we see that, for Hilbert spaces, we can replace the supremum in the definition of the norm of a linear continuous functional $\mathscr{L}$ by maximum. From now on, we use supremum and maximum interchangeably.

Example 5.5. Let $\mathbb{U} = \mathbb{R}^n$ and $\mathbb{V} = \mathbb{R}^n$ be respectively endowed with the inner products $(.,.)_{\mathbb{R}^n}$ and an $\boldsymbol{M}$-weighted inner product $(.,.)_{\mathbb{R}^n, \boldsymbol{M}}$ where $(\boldsymbol{v}, \boldsymbol{w})_{\mathbb{R}^n, \boldsymbol{M}} := \sum_{i,j} \boldsymbol{v}(i) \boldsymbol{M}(i,j) \boldsymbol{w}(j) := \boldsymbol{v}^T \boldsymbol{M} \boldsymbol{w}, \forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{V}$, and $\boldsymbol{M}$ is a symmetric and positive definite matrix. We need to find the adjoint operator $\boldsymbol{A}^*$ of a matrix $\boldsymbol{A} : \mathbb{U}, (.,.)_{\mathbb{R}^n} \to \mathbb{V}, (.,.)_{\mathbb{R}^n, \boldsymbol{M}}$. We have

$$\begin{aligned}(\boldsymbol{Au}, v)_{\mathbb{V}} &= (\boldsymbol{Au}, \boldsymbol{v})_{\mathbb{R}^n, \boldsymbol{M}} = (\boldsymbol{Av})^T \boldsymbol{Mv} \\ &= \boldsymbol{u}^T \boldsymbol{A}^T \boldsymbol{Mv} = \boldsymbol{u}^T (\boldsymbol{A}^T \boldsymbol{Mv}) = (\boldsymbol{u}, \boldsymbol{A}^T \boldsymbol{Mv})_{\mathbb{R}^n} \\ &= (\boldsymbol{u}, \boldsymbol{A}^T \boldsymbol{Mv})_{\mathbb{U}}.\end{aligned}$$

By the definition of adjoint operator, we have $\boldsymbol{A}^* = \boldsymbol{A}^T \boldsymbol{M}$.

Example 5.6. Now, let us consider $\mathscr{A} : \mathbb{U} = \text{Span}\{1,\ x,\ x^2\} \subset \mathbb{U} = \mathbb{L}^2(-1,1) \to \mathbb{R}^2$ such that the map A is defined as

$$u(x) \in \mathbb{U} \mapsto \mathscr{A}u = \begin{bmatrix} \displaystyle\int_{-1}^{1} u(x)\ dx \\ \displaystyle\int_{-1}^{1} (2x+1)u(x)\ dx \end{bmatrix},$$

with the space $\mathbb{L}^2(-1,1)$ defined[3] as

$$\mathbb{L}^2(-1,1) := \left\{ f : (-1,1) \to \mathbb{R} : \int_{-1}^{1} |f(x)|^2\ dx < \infty \right\},$$

and the inner product on $\mathbb{L}^2(-1,1)$, and hence on $\mathbb{U}$, is defined as

$$(u, v)_{\mathbb{L}^2(-1,1)} := \int_{-1}^{1} u(x) v(x)\, dx.$$

We have

$$\begin{aligned}(\mathscr{A}u,\ \boldsymbol{v})_{\mathbb{R}^2} &= \boldsymbol{v}_1 \int_{-1}^{1} u(x)\ dx + \boldsymbol{v}_2 \int_{-1}^{1} (2x+1)\ u(x)\ dx \\ &= \int_{-1}^{1} u(x) \left[\boldsymbol{v}_1 + (2x+1)\boldsymbol{v}_2\right] dx = \int_{-1}^{1} u(x) \left[1,\ (2x+1)\right] \boldsymbol{v}\ dx.\end{aligned}$$

Therefore, by definition $\mathscr{A}^* = [1,\ (2x+1)]$. It is clear that $\mathscr{A}^* \boldsymbol{v} \in \mathbb{U}$, and thus $\mathscr{A}^* : \mathbb{R}^2 \to \mathbb{U}$.

Example 5.7. Let us consider $\mathscr{A} : \mathbb{U} = \text{Span}\{1,\ x,\ x^2\} \subset \mathbb{U} = \mathbb{L}^2(-1,1) \to \mathbb{R}^2$ such that the map A is defined as

$$u(x) \in \mathbb{U} \mapsto \mathscr{A}u = \begin{bmatrix} u(x_1) \\ u(x_2) \end{bmatrix} \in \mathbb{R}^2,$$

where $x_1, x_2 \in (-1,1)$. The inner product on $\mathbb{L}^2(-1,1)$, and hence on $\mathbb{U}$, is defined as

[3] At this point, the definition of $\mathbb{L}^2(-1,1)$ is not important, but for completeness. We will see this $\mathbb{L}^2$ spaces more in latter chapters.

$$(u,v)_{\mathbb{L}^2(-1,1)} := \int_{-1}^{1} u(x)v(x)\,dx.$$

We have

$$(\mathscr{A}u,\ \boldsymbol{v})_{\mathbb{R}^2} = \boldsymbol{v}_1 u(x_1) + \boldsymbol{v}_2 u(x_2) = \int_{-1}^{1} u(x)\left[\delta(x-x_1), \delta(x-x_2)\right]\boldsymbol{v}dx.$$

Therefore, by definition $\mathscr{A}^* = [\delta(x-x_1), \delta(x-x_2)]$. Here, we have defined $\delta(x-y) \in \mathbb{U}$ via

$$\int_{-1}^{1} \delta(x-y)\,u(x)\,dx = u(y), \quad \forall u \in \mathbb{U}, \tag{5.7}$$

and thus

$$\delta(x-y) = 15\frac{\left(3y^2-1\right)}{8}x^2 + 3\frac{y}{2}x + 3\frac{\left(3-5y^2\right)}{8},$$

by testing (5.7) with $u \in \left\{1, x, x^2\right\}$. Furthermore, it is clear that $\mathscr{A}^*\boldsymbol{v} \in \mathbb{U}$, and thus $\mathscr{A}^* : \mathbb{R}^2 \to \mathbb{U}$.

Example 5.8. Consider $\mathbb{P}^n[0,1]$ the set of complex-valued polynomial of order at most n on $[0,1]$. We define $\mathscr{A} : \mathbb{U} := \mathbb{P}^n[0,1] \subset \mathbb{L}^2(0,1) \to \mathbb{U}$ as

$$\mathscr{A}u := xu' := x\frac{du}{dx},$$

and the inner product on $\mathbb{L}^2(0,1)$, and hence on $\mathbb{U}$, is defined as

$$(u,v)_{\mathbb{L}^2(0,1)} := \int_0^1 u(x)\overline{v(x)}\,dx.$$

By integration by parts we have

$$\begin{aligned}(\mathscr{A}u,v)_{\mathbb{L}^2(0,1)} = \int_0^1 xu'\overline{v}\,dx &= u(1)\overline{v(1)} + \left(u, -(xv)'\right)_{\mathbb{L}^2(0,1)} \\ &= \left(u, \delta(x-1)\,v(1) - (xv)'\right)_{\mathbb{L}^2(0,1)},\end{aligned}$$

which by definition gives

$$\mathscr{A}^*v = \delta(x-1)\,v(1) - (xv)',$$

where we have defined $\delta(x-1) \in \mathbb{U}$ via

$$\int_0^1 u(x)\overline{\delta(x-1)}\,dx = u(1), \quad \forall u \in \mathbb{U}. \tag{5.8}$$

Clearly, we can find $\delta(x-1)$ as the unique polynomial of degree at most n by testing (5.8) with $u \in \{1, x, \ldots, x^n\}$. Finally, if $v \in \mathbb{U}$, it is clear that $\mathscr{A}^* \boldsymbol{v} \in \mathbb{U}$, and thus $\mathscr{A}^* : \mathbb{U} \to \mathbb{U}$.

Proposition 5.4. *Let $\boldsymbol{u}$ and $\boldsymbol{v}$ be orthonormal bases of $\mathbb{U}$ and $\mathbb{V}$, respectively, and $\dim \mathbb{U} = n$ and $\dim \mathbb{V} = m$. Let $\boldsymbol{A}$ and $\boldsymbol{B}$ be the matrix representations of $\mathscr{A}$ and $\mathscr{A}^*$ with respect to the bases $\boldsymbol{u}$ and $\boldsymbol{v}$. Then*

$$\boldsymbol{B} = \boldsymbol{A}^*,$$

where $\boldsymbol{A}^*$ be the conjugate transpose of $\boldsymbol{A}$.

Proof. By the definition of adjoint and matrix representation, for any $u \in \mathbb{U}$ and $v \in \mathbb{V}$ and their corresponding representations $\boldsymbol{u}$ and $\boldsymbol{v}$ in the orthonormal bases $\boldsymbol{u}$ and $\boldsymbol{v}$, we have

$$(\boldsymbol{u}, \boldsymbol{B}\boldsymbol{v})_{\mathbb{F}^n} = (u, \mathscr{A}^* v)_{\mathbb{U}} = (\mathscr{A} u, v)_{\mathbb{V}} = (\boldsymbol{A}\boldsymbol{u}, \boldsymbol{v})_{\mathbb{F}^m} = (\boldsymbol{u}, \boldsymbol{A}^* \boldsymbol{v})_{\mathbb{F}^n},$$

which concludes the proof.

5.3 The closed range theorem

The following Definition 5.7, Definition 5.8, Corollary 5.1, Theorem 5.2, and Corollary 5.2 are valid for both finite and infinite dimensions.

Definition 5.7 (Orthogonal complement). Let $\mathcal{S} \subset \mathbb{U}$, the orthogonal complement $\mathcal{S}^\perp$ of $\mathcal{S}$ is defined as

$$\mathcal{S}^\perp := \{u \in \mathbb{U} : (u, w)_{\mathbb{U}} = 0, \forall w \in \mathcal{S}\}.$$

A direct consequence of the definition is that $\mathcal{S}^\perp$ is a closed subspace of $\mathbb{U}$ and that $\mathcal{S} \cap \mathcal{S}^\perp = \{\theta\}$.

Definition 5.8 (Closure). Let $\mathcal{S} \in \mathbb{U}$. The closure $\overline{\mathcal{S}}$ of $\mathcal{S}$ is the smallest closed set in $\mathbb{U}$ containing $\mathcal{S}$.

Note that we use the overline to denote both the complex conjugate and the closure, but it should be clear from the context which one we refer to.

Remark 5.6. Note that the closure $\overline{\mathcal{S}}$ is the smallest closed set containing $\mathcal{S}$, and thus the closure $\overline{\mathcal{S}}$ is unique. In addition, iof $\mathcal{S}$ is a linear subspace, then the closure under the $\mathbb{U}$-norm topology is the same as the completion of $\mathcal{S}$ with the $\mathbb{U}$-norm. As a result, the completion of a linear subspace in a normed space is unique.

Corollary 5.1. *There holds:* $\left(\mathcal{S}^{\perp}\right)^{\perp} = \overline{\mathcal{S}}$.

Proof. See [97, proposition 1 of chapter 3].

Next is an important theorem (see, e.g., [141, 28, 14]).

Theorem 5.2 (The closed range theorem). *Let* $\mathscr{A} : \mathbb{U} \to \mathbb{V}$. *The following hold:*

- $[\mathsf{R}(\mathscr{A})]^{\perp} = \mathsf{N}(\mathscr{A}^*)$.
- $\overline{\mathsf{R}(\mathscr{A})} = [\mathsf{N}(\mathscr{A}^*)]^{\perp}$.
- $[\mathsf{R}(\mathscr{A}^*)]^{\perp} = \mathsf{N}(\mathscr{A})$.
- $\overline{\mathsf{R}(\mathscr{A}^*)} = [\mathsf{N}(\mathscr{A})]^{\perp}$.[4]

If $\mathsf{R}(\mathscr{A})$ *is closed, so is* $\mathsf{R}(\mathscr{A}^*)$, *and we can replace* $\overline{\mathsf{R}(\mathscr{A})}$ *and* $\overline{\mathsf{R}(\mathscr{A})^*}$ *by* $\mathsf{R}(\mathscr{A})$ *and* $\mathsf{R}(\mathscr{A}^*)$, *respectively, in the above results.*

Proof. The second assertion is the direct consequence of the first assertion and Corollary 5.1. The third and fourth assertions follow the first and the second, and the fact that $(\mathscr{A}^*)^* = \mathscr{A}$. So, we only need to prove the first assertion. Let $z \in \mathsf{N}(\mathscr{A}^*)$ and $y \in \mathsf{R}(\mathscr{A})$. Then $y = \mathscr{A}x$ for some $x \in \mathbb{U}$. We have

$$(z, y)_{\mathbb{V}} = (z, \mathscr{A}x)_{\mathbb{V}} = (\mathscr{A}^* z, x)_{\mathbb{U}} = 0, \quad \forall y \in \mathsf{R}(\mathscr{A}),$$

which says that $\mathsf{N}(\mathscr{A}^*) \subset [\mathsf{R}(\mathscr{A})]^{\perp}$. Now take $z \in [\mathsf{R}(\mathscr{A})]^{\perp}$, we have

$$(\mathscr{A}^* z, x)_{\mathbb{U}} = (z, \mathscr{A}x)_{\mathbb{V}} = 0, \quad \forall x \in \mathbb{U},$$

which implies that $\mathscr{A}^* z = 0$, which in turn shows $[\mathsf{R}(\mathscr{A})]^{\perp} \subset \mathsf{N}(\mathscr{A}^*)$.

Corollary 5.2. *There hold:*

- $\mathbb{U} = \mathsf{N}(\mathscr{A}) \oplus \overline{\mathsf{R}(\mathscr{A}^*)}$,
- $\mathbb{V} = \mathsf{N}(\mathscr{A}^*) \oplus \overline{\mathsf{R}(\mathscr{A})}$,

where $\oplus$ *denotes the direct sum of two sets, that is, for example, for any* $u \in \mathbb{U}$ *there are unique* $u_1 \in \mathsf{N}(\mathscr{A})$ *and* $u_2 \in \overline{\mathsf{R}(\mathscr{A}^*)}$ *such that* $u = u_1 + u_2$.

We note that for finite dimensional vector spaces $\mathbb{U}$ and $\mathbb{V}$, $\mathsf{R}(\mathscr{A})$, and hence $\mathsf{R}(\mathscr{A}^*)$, is obviously closed (see Problem 5.9), and thus Theorem 5.2 and Corollary 5.2 clearly hold with $\mathsf{R}(\mathscr{A})$ and $\mathsf{R}(\mathscr{A}^*)$ in the places of $\overline{\mathsf{R}(\mathscr{A})}$ and $\overline{\mathsf{R}(\mathscr{A}^*)}$. We will see that the proof of the closed range Theorem 5.2, hence the Corollary 5.2, for finite dimensional cases is trivial using the SVD decomposition in Theorem 8.1.

When $\mathscr{A}$ is continuous and its the range space is finite-dimensional (also called finite-rank), we can say a lot about $\mathscr{A}$ using its adjoint $\mathscr{A}^*$.

[4] Since we consider only Hilbert spaces, which are reflexive, $\overline{\mathsf{R}(\mathscr{A}^*)} = [\mathsf{N}(\mathscr{A})]^{\perp}$ holds. In general, $\overline{\mathsf{R}(\mathscr{A}^*)} \subset [\mathsf{N}(\mathscr{A})]^{\perp}$: see [28, Corollary 2.18].

Lemma 5.2 (Finite-rank operators). *Let $\mathscr{A} \in \mathcal{B}(\mathbb{U}, \mathbb{V})$ and suppose $\dim \mathsf{R}(\mathscr{A}) = n < \infty$. Let $\{\varphi^i\}_{i=1}^n$ be an orthonormal basis of $\mathsf{R}(\mathscr{A})$. Then,*

- *$\mathscr{A}$ can be expressed as*

$$\mathscr{A}u = \sum_{i=1}^{n} \left(\phi^i, u\right)_{\mathbb{U}} \varphi^i,$$

  where $\{\phi^i\}_{i=1}^n$ be a linearly independent set in $\mathbb{U}$.
- *$\mathscr{A}^* \in \mathcal{B}(\mathbb{V}, \mathbb{U})$ and can be expressed as*

$$\mathscr{A}^* v = \sum_{i=1}^{n} \left(\varphi^i, v\right)_{\mathbb{V}} \phi_i$$

- $\dim(\mathsf{R}(\mathscr{A}^*)) = \dim(\mathsf{R}(\mathscr{A})) = n$.

Proof. There exists a unique coordinate vector $\boldsymbol{\alpha} \in \mathbb{F}^n$ such that

$$\mathscr{A}u = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \varphi^i,$$

which, after taking the inner product both sides with φ^j, reduces to

$$\overline{\boldsymbol{\alpha}_j} = \left(\mathscr{A}u, \varphi^j\right)_{\mathbb{V}} = \left(u, \mathscr{A}^* \varphi^j\right)_{\mathbb{U}} = \left(u, \phi^j\right)_{\mathbb{U}},$$

where we have defined $\phi^j := \mathscr{A}^* \varphi^j$. Note that the computation of $\boldsymbol{\alpha}_j, j = 1, \ldots, n$, is well-defined due to the boundedness of $\mathscr{A}$. The fact that $\{\phi^j\}_{j=1}^n$ is an independent set of $\mathbb{U}$ is due to the independence of $\{\varphi^i\}_{i=1}^n$. For the second assertion, the characterization of $\mathscr{A}^*$ is obvious using the definition of adjoint, and the boundedness of $\mathscr{A}^*$ then follows. The third assertion is clear from the first and the second assertion.

Theorem 5.3 (Rank-Nullity Theorem). *Let $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ is a linear operator and $\dim(\mathbb{U}), \dim(\mathbb{V}) < \infty$. The following hold true*

- $\dim(\mathbb{U}) = \dim(\mathsf{N}(\mathscr{A})) + \dim(\mathsf{R}(\mathscr{A}))$,
- $\dim(\mathbb{V}) = \dim(\mathsf{N}(\mathscr{A}^*)) + \dim(\mathsf{R}(\mathscr{A}^*))$.

Proof. By Corollary 5.2 and finite dimensional nature of $\mathbb{U}$, we know that $\mathbb{U} = \mathsf{N}(\mathscr{A}) \oplus \overline{\mathsf{R}(\mathscr{A}^*)}$, and thus

$$\dim(\mathbb{U}) = \dim(\mathsf{N}(\mathscr{A})) + \dim(\mathsf{R}(\mathscr{A}^*)),$$

which, together with the third assertion of Lemma 5.2, yields the first assertion. The proof of the second assertion follows similarly. An alternative and straightforward proof using the SVD can be referred to section 8.1.

5.4 Appendix: an origin of matrices and their rules

This section review the notion of matrix representation of a linear mapping between two finite dimensional spaces. Our goal is to provide a constructive derivation of matrices and vectors, and their operational rules. As shall be seen, these are a by-product of viewing a linear transformation under a particular basis in the domain and a particular basis in the range of a linear map. Consider the linear operator: $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ where $dim(\mathbb{U}) = n,\ dim(\mathbb{V}) = m$. We find a matrix representation of $\mathscr{A}$ by the following steps:

1. Let
$$v := \mathscr{A}u \tag{5.9}$$
Let $\boldsymbol{E} = \left\{e^1, \ldots, e^n\right\}$ is the Hamel basis of $\mathbb{U}$ and $\boldsymbol{G} = \left\{g^1, \ldots, g^m\right\}$ is the Hamel basis of $\mathbb{V}$. Now, any $v := \mathscr{A}u \in \mathbb{V}$ can be written as
$$v = Au = \mathscr{A}\left(\sum_{j=1}^{n} \boldsymbol{\alpha}_j e^j\right) = \sum_{j=1}^{n} \boldsymbol{\alpha}_j \mathscr{A}e^j,$$
Note that $[\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_n]^T$ are the coordinates of u in the basis $\mathbb{X}$. Since $\mathscr{A}e^j \in V$, there exists a unique m numbers $\widehat{\boldsymbol{A}}_{ij}, i = 1 \ldots m$ such that
$$\mathscr{A}e^j = \sum_{i=1}^{m} \widehat{\boldsymbol{A}}_{ij} g^i.$$
Thus,
$$v = \sum_{j=1}^{n} \boldsymbol{\alpha}_j \mathscr{A}e^j = \sum_{j=1}^{n} \boldsymbol{\alpha}_j \left(\sum_{i=1}^{m} \widehat{\boldsymbol{A}}_{ij} g^i\right).$$
Switching the sum gives
$$v = \sum_{i=1}^{m} \left(\sum_{j=1}^{n} \widehat{\boldsymbol{A}}_{ij} \boldsymbol{\alpha}_j\right) g^i.$$
On the other hand, since $v \in V$, there exists a unique set of coordinates $\{\boldsymbol{\beta}_i\}_{i=1}^m$ of v in the basis $\boldsymbol{G}$ such that
$$v = \sum_{i=1}^{m} \boldsymbol{\beta}_i g^i.$$
Combining the last two equations yields

$$\sum_{i=1}^{m}\left(\boldsymbol{\beta}_i-\sum_{j=1}^{n}\widehat{\boldsymbol{A}}_{ij}\boldsymbol{\alpha}_j\right)g^i=0, \tag{5.10}$$

which, owing to the linear dependence of $\boldsymbol{G}$, gives

$$\boldsymbol{\beta}_i=\sum_{j=1}^{n}\widehat{\boldsymbol{A}}_{ij}\boldsymbol{\alpha}_j,\quad i=1,\ldots,m. \tag{5.11}$$

2. If we put $\{\boldsymbol{\alpha}_j\}_{j=1}^n$, $\{\boldsymbol{\beta}_i\}_{i=1}^m$, and $\left\{\widehat{\boldsymbol{A}}_{ij}\right\}_{i=1,j=1}^{n,m}$ into tables/arrays of numbers as

$$\underbrace{\begin{Bmatrix}\boldsymbol{\beta}_1\\ \vdots\\ \boldsymbol{\beta}_m\end{Bmatrix}}_{\boldsymbol{\beta}:=},\quad \underbrace{\begin{bmatrix}\widehat{\boldsymbol{A}}_{11} & \widehat{\boldsymbol{A}}_{12} & \ldots & \widehat{\boldsymbol{A}}_{1n}\\ \vdots & \ddots & \ddots & \vdots\\ \widehat{\boldsymbol{A}}_{m1} & \ldots & \ldots & \widehat{\boldsymbol{A}}_{mn}\end{bmatrix}}_{\widehat{\boldsymbol{A}}:=},\quad \underbrace{\begin{Bmatrix}\boldsymbol{\alpha}_1\\ \vdots\\ \boldsymbol{\alpha}_n\end{Bmatrix}}_{\boldsymbol{\alpha}:=},$$

then (5.11) can be be written in a short-hand notation as

$$\boldsymbol{\beta}=\boldsymbol{A}\times\boldsymbol{\alpha},$$

where the multiplication $\times$ between $\boldsymbol{A}$ and $\boldsymbol{\alpha}$ is defined via (5.11). In particular, the ith component of $\boldsymbol{\beta}$ is computed by the Euclidean inner product of the ith row of $\boldsymbol{A}$ and $\boldsymbol{\alpha}$. We call $\boldsymbol{A}$ an $m\times n$ matrix, $\boldsymbol{\alpha}$ an n-dimensional vector, and $\boldsymbol{\beta}$ an m-dimensional vector. *Thus, we have derived the product rule between a matrix and a vector that we were taught to memorize.* We should have been taught the origin of matrix multiplication and addition rules as a direct consequence of matrix representations of a linear map and the sum of two linear maps. This provides not only the origin of matrices/vectors, but also the underlying linear map behind a matrix. Such an introduction provides a deeper understanding of matrices. Unfortunately, most linear algebra books/courses regarding matrices provide the rules out of the blue. As we have seen, deriving these rules is trivial, and yet it provides insights into the origin of matrices and their rules. Other matrix rules can be derived in a similar fashion (see the derivation of matrix addition and matrix multiplication rules in Problem 5.14 and Problem 5.15).
3. $\widehat{\boldsymbol{A}}$ is known as the matrix representation of the linear operator $\mathscr{A}$ in the $(\boldsymbol{E},\boldsymbol{G})$ basis pair. As can be seen, $\boldsymbol{\alpha}$ ($\boldsymbol{\beta}$ respectively) is the presentation of u in the basis $\boldsymbol{E}$ ($\boldsymbol{G}$ respectively).
4. Note that the original linear map $\mathscr{A}$ is completely equivalent its matrix presentation $\boldsymbol{A}$. While the matrix representation $\boldsymbol{A}$ depends on the pair of bases under consideration, the original operator is independent of bases. Similarly, u (and v respectively) is basis-independent, its representation $\boldsymbol{\alpha}$ (and $\boldsymbol{\beta}$ respectively) is basis-dependent.

Problems

Problem 5.1. Let $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ be a linear operator. Show that $\mathsf{R}(\mathscr{A})$ and $\mathsf{N}(\mathscr{A})$ are vector spaces.

Problem 5.2. Prove the statement in Remark 5.2.

Problem 5.3. Recall that if $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ is bijective (injective and surjective), its inverse $\mathscr{A}^{-1} : \mathbb{V} \to \mathbb{U}$ is well-defined in the sense that for any $v \in \mathbb{V}$, there is a unique $u \in \mathbb{U}$ such that $\mathscr{A}u = v$ and $\mathscr{A}^{-1}(v) = u$. Show that $\mathscr{A}^{-1}$ is also a linear operator.

Problem 5.4. Prove Theorem 5.1 for the case when $\mathbb{F} = \mathbb{R}$.

Problem 5.5. Let $\mathscr{A} \in \mathcal{B}(\mathbb{U}, \mathbb{V})$, $\mathscr{I} : \mathbb{U} \to \mathbb{U}$ is the identity operator, and $\lambda \in \mathbb{C}$. Prove the following facts:

- $\mathscr{I}^* = \mathscr{I}$,
- $(\lambda\mathscr{A})^* = \overline{\lambda}\mathscr{A}^*$,
- $(\mathscr{A} + \lambda\mathscr{I})^* = \mathscr{A}^* + \overline{\lambda}\mathscr{I}$,
- If $\mathscr{A}^{-1}$ exists, then $(\mathscr{A}^*)^{-1} = \left(\mathscr{A}^{-1}\right)^*$,
- Suppose $\mathscr{B} \in \mathcal{B}(\mathbb{V}, \mathbb{Z})$, then $(\mathscr{B}\mathscr{A})^* = \mathscr{A}^*\mathscr{B}^*$.

Problem 5.6. Let $\mathbb{U} = \mathbb{C}^n$ and $\mathbb{V} = \mathbb{C}^n$ be respectively endowed with the an $\boldsymbol{N}$-weighted $(.,.)_{\mathbb{C}^n,\boldsymbol{N}}$ and an $\boldsymbol{M}$-weighted inner product $(.,.)_{\mathbb{C}^n,\boldsymbol{M}}$ where, for example, $(\boldsymbol{v},\boldsymbol{w})_{\mathbb{C}^n,\boldsymbol{M}} := \sum_{i,j} \boldsymbol{v}(i)\boldsymbol{M}(i,j)\overline{\boldsymbol{w}(j)} := \boldsymbol{v}^T\boldsymbol{M}\overline{\boldsymbol{w}}, \forall \boldsymbol{v},\boldsymbol{w} \in \mathbb{V}$. Here, $\boldsymbol{N}$ and $\boldsymbol{M}$ are a Hermitian, i.e. $\boldsymbol{M}^* = \boldsymbol{M}$, and positive definite matrices. Find the adjoint operator $\boldsymbol{A}^*$ of a matrix $\boldsymbol{A} : \mathbb{U}, (.,.)_{\mathbb{C}^n,\boldsymbol{N}} \to \mathbb{V}, (.,.)_{\mathbb{C}^n,\boldsymbol{M}}$.

Problem 5.7. Now, let us consider $\mathscr{A} : \mathbb{U} = \text{Span}\{1,\ x,\ x^2\} \subset \mathbb{U} = \mathbb{L}^2(-1,1) \to \mathbb{R}^2$ such that the map A is defined as

$$u(x) \in \mathbb{U} \mapsto \mathscr{A}u = \begin{bmatrix} \int_{-1}^{1} u(x)\ dx \\ u(x_0) \end{bmatrix},$$

where $u_0 \in (-1,1)$ and the inner product on $\mathbb{L}^2(-1,1)$, and hence on $\mathbb{U}$, is defined as

$$(u,v)_{\mathbb{L}^2(-1,1)} := \int_{-1}^{1} w(x)\,u(x)v(x)\,dx,$$

where $w(x)$ is a given positive function $(-1,1)$. Find the adjoint of $\mathscr{A}$.

Problem 5.8. Consider $\mathbb{P}^n[0,1]$ the set of complex-valued polynomial of order at most n on $[0,1]$. We define $\mathscr{A} : \mathbb{U} := \mathbb{P}^n[0,1] \subset \mathbb{L}^2(0,1) \to \mathbb{U}$ as

$$\mathscr{A}u := \left(u^2+1\right)u' := \left(u^2+1\right)\frac{du}{dx},$$

and the inner product on $\mathbb{L}^2(0,1)$, and hence on $\mathbb{U}$, is defined as

$$(u,v)_{\mathbb{L}^2(0,1)} := \int_0^1 \overline{u(x)} v(x)\, dx.$$

Problem 5.9. Let $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ be a continuous linear map. Suppose that either $\mathbb{U}$ or $\mathbb{V}$ is a finite-dimensional space. Show that $\mathsf{R}(\mathscr{A})$ is finite-dimensional space. Then, show that $\mathsf{R}(\mathscr{A})$, and thus $\mathsf{R}(\mathscr{A}^*)$, is closed.

Hint: Suppose that $\dim(\mathbb{U}) = n < \infty$. Let $u_1, \ldots, u_n$ be an orthonormal basis for $\mathbb{U}$. Then clearly $\mathscr{A}u_i$, $i = 1, \ldots, n$, spans $\mathsf{R}(\mathscr{A})$, and thus $\dim(\mathsf{R}(\mathscr{A})) \leq n$. The equality holds when $\mathsf{N}(\mathscr{A}) = \{\theta\}$, which we can assume WLOG. Now takes a convergent sequence $\{v_i\}_{i=1}^{\infty}$ in $\mathsf{R}(\mathscr{A})$ that converges to v, then the pre-image sequence $\{e_i\}_{i=1}^{\infty}$ is (Cauchy and thus) convergent as $\mathscr{A}$ is bounded and invertible on its range. Let u be the limit of $\{e_i\}_{i=1}^{\infty}$, then $v = \lim_{i\to\infty} y_i = \lim_{i\to\infty} \mathscr{A}u_i = \mathscr{A}u$. This shows that $v \in \mathsf{R}(\mathscr{A})$ and this concludes the proof.

Note that when $\dim \mathbb{U} < \infty$, $\mathscr{A}$ is automatically continuous (see Proposition 5.1)

When $\dim(\mathbb{V}) = n < \infty$, we do need the continuity assumption of $\mathscr{A}$ to show that it is a compact operator. Then we can show that $\mathsf{R}(\mathscr{A})$ is closed iff $\mathsf{R}(\mathscr{A})$ is finite dimensional. **But this second part may be too much: perhaps should eliminate this second part and only consider the case** $\dim \mathbb{U} < \infty$.

Problem 5.10. Let $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ be a continuous linear map and $\dim(\mathsf{R}\{\mathscr{A}\}) = n < \infty$. Let $\{v_i\}_{i=1}^{n}$ be an orthonormal basis of $\mathsf{R}(\mathscr{A})$. Show that $\mathscr{A}$ can be expressed as

$$\mathscr{A}u = \sum_{i=1}^{n} (u_i, u)_{\mathbb{U}}\, v_i,$$

where $\{u_i\}_{i=1}^{n}$ be a linearly independent set in $\mathbb{U}$. In addition, shows that $\dim(\mathsf{R}(\mathscr{A}^*)) = \dim(\mathsf{R}(\mathscr{A})) = n$. Conversely, if $\dim(\mathsf{R}\{\mathscr{A}^*\}) = n < \infty$, then $\dim(\mathsf{R}(\mathscr{A})) = \dim(\mathsf{R}(\mathscr{A}^*)) = n$.

Hint: Let $\mathscr{A}u = \sum_{i=1}^{n} \alpha_i v_i$. By inner product both sides by v_j, we see that $\overline{\alpha_i} = (\mathscr{A}u, v_i)_{\mathbb{V}} = (u, \mathscr{A}^* v_i)_{\mathbb{U}}$. Note that due to the continuity (boundedness) of $\mathscr{A}$, the computation of α_i makes sense for any $u \in \mathbb{U}$. Otherwise, α_i could be infinite if $\mathscr{A}$ is unbounded. By setting $u_i = \mathscr{A}^* v_i$ concludes the result. Consequently, we have

$$\mathscr{A}^* v = \sum_{i=1}^{n} (v_i, v)_{\mathbb{V}}\, u_i,$$

Problem 5.11. Let us equip $\mathcal{L}(\mathbb{U}, \mathbb{V})$ (and $\mathcal{B}(\mathbb{U}, \mathbb{V})$) with the following algebraic operations for $f, g \in \mathcal{L}(\mathbb{U}, \mathbb{V})$ (and $\mathcal{B}(\mathbb{U}, \mathbb{V})$):

$$(f+g)(u) := f(u) + g(u),$$
$$(\alpha f)(u) := \alpha f(u), \quad \forall \alpha \in \mathbb{F}.$$

Show that $\mathcal{L}(\mathbb{U}, \mathbb{V})$ (and $\mathcal{B}(\mathbb{U}, \mathbb{V})$) is a vector space.

Problem 5.12. Consider $u \in \mathbb{U}$ and define the Riesz map $\mathcal{R}_u$ via its action $\mathbb{U} \ni v \mapsto \mathcal{R}_u(v) := \langle \mathcal{R}_u, v\rangle := (u, v)_{\mathbb{U}} \in \mathbb{F}$. The linearity of the Riesz map $\mathcal{R}$ is clear, and thus we write $\mathcal{R}u$ instead of $\mathcal{R}_u$. Show that

- $\mathcal{R}u \in \mathbb{U}^*$ for any $u \in \mathbb{U}$,
- $\|\mathcal{R}u\| = \|u\|_{\mathbb{U}}$ for any $u \in \mathbb{U}$ (and thus $\mathcal{R}$ is an isometry), and
- $\mathcal{R}(\cdot) : \mathbb{U} \ni u \mapsto \mathcal{R}u \in \mathbb{U}^*$ is injective.

Together with the Riesz representation Theorem 5.1 and adjoint Definition 5.6, show that $\mathcal{R}$ is a unitary operator.

Hint: The first and the second assertions are clear by the Cauchy-Schwarz inequality. For the third assertion, suppose $\mathcal{R}u = \mathcal{R}w$ and $w \neq u$. Then $(w, v)_{\mathbb{U}} = \mathcal{R}_w(v) = \mathcal{R}_u(v) = (u, v)_{\mathbb{U}}, \forall v \in \mathbb{U}$, and thus $(u - w, v)_{\mathbb{U}}, \forall v \in \mathbb{U}$. Then simply taking $v = u - w$, we have that $u = w$. For the last assertion, we have

$$\|u\|^2 = \|\mathcal{R}u\|^2 = (\mathcal{R}^*\mathcal{R}u, u), \implies ((\mathcal{R}^*\mathcal{R} - \mathscr{I})u, u) = 0 \quad \forall u \in \mathbb{U},$$

which, by taking $u \mapsto u + w$ and $u \mapsto u + iw$ for examples, implies that $\mathcal{R}^*\mathcal{R} = \mathscr{I}$. From the Riesz representation theorem, $\mathcal{R}$ is subjective, and thus for any $\varphi \in \mathbb{U}^*$, there exists a unique $v \in \mathbb{U}$ such that $\mathcal{R}v = \varphi$, and thus

$$\mathcal{R}\mathcal{R}^*\varphi = \mathcal{R}\mathcal{R}^*\mathcal{R}v = \mathcal{R}v = \varphi, \quad \forall \varphi \in \mathbb{U}^*,$$

where have used the fact that $\mathcal{R}^*\mathcal{R} = \mathscr{I}$ in the second last equality. Thus $\mathcal{R}\mathcal{R}^* = \mathscr{I}$. We conclude that $\mathcal{R} : \mathbb{U} \to \mathbb{U}^*$ is indeed a unitary operator.

Problem 5.13. Show that $\mathbb{U}^*$ equipped with the inner product in (5.5) is a Hilbert space, and the induced norm is exactly the operator norm in $\mathbb{U}^*$.

Problem 5.14 (matrix addition). Consider $\mathscr{A}, \mathscr{B} : \mathbb{U} \to \mathbb{V}$ with $\dim \mathbb{U} = n$ and $\dim \mathbb{V} = m$. Let $\boldsymbol{E}$ and $\boldsymbol{G}$ be a basis for $\mathbb{U}$ and $\mathbb{V}$, respectively, and the corresponding matrix representations for $\mathscr{A}$ and $\mathscr{B}$ in this basis pair are $\boldsymbol{A}$ and $\boldsymbol{B}$. Let $\boldsymbol{C}$ be the matrix presentation of $\mathscr{C} := \mathscr{A} + \mathscr{B} : \mathbb{U} \to \mathbb{V}$. Show that

$$\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B},$$

where the addition operation is componentwise.

Problem 5.15 (matrix multiplication). Consider $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ and $\mathscr{B} : \mathbb{V} \to \mathbb{U}$ with $\dim \mathbb{U} = n$ and $\dim \mathbb{V} = m$. Let $\boldsymbol{E}$ and $\boldsymbol{G}$ be a basis for $\mathbb{U}$ and $\mathbb{V}$, respectively, and the corresponding matrix representations for $\mathscr{A}$ and $\mathscr{B}$ in this basis pair are $\boldsymbol{A}$ and $\boldsymbol{B}$. Let $\boldsymbol{C}$ be the matrix presentation of the composition $\mathscr{C} := \mathscr{B}\mathscr{A} : \mathbb{U} \to \mathbb{U}$. Show that

$$\boldsymbol{C} = \boldsymbol{B} \times \boldsymbol{A},$$

where the multiplication produces the ij-component of matrix $\boldsymbol{C}$ as follow

$$\boldsymbol{C}_{ij} = \sum_{k=1}^{m} \boldsymbol{B}_{ik} \boldsymbol{A}_{kj} = \left(\boldsymbol{B}\left(i,:\right), \boldsymbol{A}\left(:,j\right) \right)_{\mathbb{R}^m},$$

where $\boldsymbol{B}\left(i,:\right)$ and $\boldsymbol{A}\left(:,j\right)$ are ith row and jth column of $\boldsymbol{B}$ and $\boldsymbol{A}$, respectively.

Chapter 6
Existence and uniqueness of a solution for linear operator equations

Abstract In this chapter, we study the existence and uniqueness of a solution of the following linear operator equation

$$\mathscr{A}u = f, \tag{6.1}$$

where $\mathscr{A} : \mathbb{U} \to \mathbb{V}$. This short chapter is limited to finite dimensional spaces $\mathbb{U}$ and $\mathbb{V}$, while addressing similar questions in arbitrary dimensions, including infinite dimensions, will be the main subject of Chapter 15. We shall see that the adjoint plays a central role in the existence and uniqueness of a solution for (6.1). The prerequisites for this chapter are:

- Linear operator (basic definitions including range and null spaces: see Chapter 5)
- Basics on Hilbert spaces (inner product, orthogonality)
- Linear algebra

By definition, there exists a solution to (6.1) iff $f \in \mathsf{R}(\mathscr{A})$. Thus, one way to determine the existence of a solution is by studying the range space $\mathsf{R}(\mathscr{A})$ of $\mathscr{A}$. However, there are two possible issues here. First, it may not be trivial to identify $\mathsf{R}(\mathscr{A})$. Second, even when we know $\mathsf{R}(\mathscr{A})$, there is no simple working mechanics to check if f resides in $\mathsf{R}(\mathscr{A})$. Fortunately, the closed range Theorem 5.2, tells us that $\mathsf{R}(\mathscr{A}) = \mathsf{N}(\mathscr{A}^*)^\perp$ as long as $\mathsf{R}(\mathscr{A})$ is closed, which is the case for this chapter as $\dim \mathbb{X} = n < \infty$. Thus, the existence of a solution now amounts to verifying that $f \in \mathsf{N}(\mathscr{A}^*)^\perp$, which is equivalent to checking

$$(f, v)_{\mathbb{Y}} = 0, \quad \forall v \in \mathsf{N}(\mathscr{A}^*),$$

and this provides us with a working equation to operate on. Of course, instead of characterizing $\mathsf{R}(\mathscr{A})$ we need to identify $\mathsf{N}(\mathscr{A}^*)$, but this can be done through the following working equation

$$\mathscr{A}^* y = \theta, \quad \forall y \in \mathsf{N}(\mathscr{A}^*).$$

Clearly, the additional step that we have to carry out is to identify $\mathscr{A}^*$. We will demonstrate these steps in details for various examples in the following.

Lemma 6.1.

- **Existence.** *The linear equation (6.1) has a solution iff* $y \in \mathsf{N}(\mathscr{A}^*)^\perp$.
- **Uniqueness.** *The solution of (6.1) is unique iff* $[\mathsf{R}(\mathscr{A}^*)]^\perp = \mathsf{N}(\mathscr{A}) = \{\theta\}$.
- *If* $\dim \mathbb{X} = \dim \mathbb{Y}$, *the uniqueness is equivalent to existence.*

Proof. As discussed above, the existence is the direct consequence of the closed range Theorem 5.2 which says that $\mathsf{R}(\mathscr{A}) = [\mathsf{N}(\mathscr{A}^*)]^\perp$. The uniqueness is clear. The proof of the third assertion is as follows. We have

$$\begin{aligned}
\mathsf{N}(\mathscr{A}) &= \{\theta\} && \text{Uniqueness}\\
&\Updownarrow\\
\dim \mathsf{N}(\mathscr{A}) &= 0\\
&\Updownarrow && \dim \mathbb{X} = \dim \mathsf{N}(\mathscr{A}) + \dim \mathsf{R}(\mathscr{A})\\
\dim \mathbb{Y} = \dim \mathbb{X} &= \dim \mathsf{R}(\mathscr{A})\\
&\Updownarrow && \mathsf{R}(\mathscr{A}) \subseteq \mathbb{Y}\\
\mathsf{R}(\mathscr{A}) &= \mathbb{Y} && \text{Existence for any } y \in \mathbb{Y}
\end{aligned}$$

where we have used the rank-nullity Theorem 5.3.

Remark 6.1. Lemma 6.1 implies that if a solution exists, the natural approach to study its uniqueness is to investigate the null space of $\mathscr{A}$. However, there are cases where we can readily see the answer from the range of $\mathscr{A}^*$.

Example 6.1. Let $\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$, and $\boldsymbol{f} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The question is if there is a solution to the equation $\boldsymbol{Au} = \boldsymbol{f}$. Consider $\mathbb{U} = \mathbb{V} = \mathbb{R}^2$ with the standard Euclidean inner product $(\cdot, \cdot)_{\mathbb{R}^2}$. We know from Example 5.5 that the adjoint $\boldsymbol{A}^*$ is given by

$$\boldsymbol{A}^* = \boldsymbol{A}^T = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}.$$

Let's determine the null space of $\boldsymbol{A}^*$. If $\mathbf{z} \in \mathsf{N}(\boldsymbol{A}^*)$ then

$$\boldsymbol{A}^*\mathbf{z} = \theta \implies \boldsymbol{z}_1 + \boldsymbol{z}_2 = 0,$$

which yields $\mathsf{N}(\boldsymbol{A}^*) = \left\{ \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix} : \quad \forall \alpha \in \mathbb{R} \right\}$. However, for any $\boldsymbol{z} \in \mathsf{N}(\boldsymbol{A}^*)$ we have

$$(\boldsymbol{f}, \boldsymbol{z})_{\mathbb{R}^2} = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix} \right)_{\mathbb{R}^2} = \alpha \quad \forall \alpha \in \mathbb{R},$$

that is, $\boldsymbol{f} \not\perp \mathsf{N}(\boldsymbol{A}^*)$. Thus, $\boldsymbol{Au} = \boldsymbol{f}$ does not have a solution.

Example 6.2. Consider the operator $\mathscr{A}$ defined in Example 5.6: $\mathscr{A} : \mathbb{U} = \text{Span}\{1,\ x,\ x^2\} \subset \mathbb{U} = \mathbb{L}^2(-1,1) \to \mathbb{R}^2$ such that

$$u(x) \in \mathbb{U} \mapsto \mathscr{A}u = \begin{bmatrix} \int_{-1}^{1} u(x)\ dx \\ \int_{-1}^{1} (2x+1)u(x)\ dx \end{bmatrix},$$

and we are interested in studying the existence and uniqueness of a solution for $\mathscr{A}u = \boldsymbol{f}$ for a given $\boldsymbol{f} \in \mathbb{R}^2$. Note that this is an operator setting for the problem of fitting a quadratic polynomial in $u(x)$ with two pieces of information about $u(x)$. With elementary knowledge, we know that there is more than one quadratic as we can only hope to determine a quadratic uniquely with three pieces of information. We expect the proposed adjoint approach to provide the same answer. To that end, we see that $\mathscr{A}$ maps three-dimensional space $\mathbb{U}$ into two-dimensional space $\mathbb{R}^2$, but other than this it is unclear what the range and null spaces of $\mathscr{A}$ look like. However, the range and null spaces of $\mathscr{A}^*$ are trivial as we shall see. We start by recalling from Example 5.6 that

$$\mathscr{A}^* = [1,\ (2x+1)]\,.$$

To compute $\mathsf{N}(\mathscr{A}^*)$, we pick any $v \in \mathsf{N}(\mathscr{A}^*)$, i.e. $\mathscr{A}^*\boldsymbol{v} = 0$. We have

$$\boldsymbol{v}_1 + (2x+1)\boldsymbol{v}_2 = 0 \quad \forall x,$$

which implies

$$\boldsymbol{v}_1 = \boldsymbol{v}_2 = 0.$$

Thus,

$$\mathsf{N}(\mathscr{A}^*) = \left\{[0,0]^T\right\}.$$

Since $[0,0]^T$ is orthogonal to any $\boldsymbol{f} \in \mathbb{R}^2$, we conclude that $\mathscr{A}u = \boldsymbol{f}$ always has a solution.

Now for the uniqueness, we have seen that $\mathscr{A}^*\boldsymbol{v} = \boldsymbol{v}_1 + (2x+1)\boldsymbol{v}_2$ which is a first-order polynomial while $\mathbb{U}$ is the space of polynomials of order at most 2. It follows that $\mathsf{R}(\mathscr{A}^*) \subset \mathbb{U}$, which means $\mathsf{N}(\mathscr{A})$, the orthogonal complement of $\mathsf{R}(\mathscr{A}^*)$ in $\mathbb{U}$ is not trivial. Thus, there is no uniqueness. To see that this indirect approach via adjoint yields the same conclusion using the direction approach based on $\mathsf{N}(\mathscr{A})$, let us now find $\mathsf{N}(\mathscr{A})$.

To compute $\mathsf{N}(\mathscr{A})$, we pick any $u \in \mathsf{N}(\mathscr{A})$, i.e. $\mathscr{A}u = [0,0]^T$, i.e:

$$\begin{bmatrix} \int_{-1}^{1} u\ dx \\ \int_{-1}^{1} (2x+1)\ u\ dx \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

Since $\mathbb{U} = \text{Span}\{1,\ x,\ x^2\}$, we can write $u(x) = ax^2 + bx + c$ where $a, b, c \in \mathbb{R}$ need to be determined. Substituting into the above equations yields:

$$b = 0 \quad \text{and} \quad a = -3c.$$

That is,

$$u(x) = (-3x^2 + 1)c,$$

for any constant $c \in \mathbb{R}$. We conclude

$$\mathsf{N}(\mathscr{A}) = \{(-3x^2 + 1)c :\ c \in \mathbb{R}\}$$

is not trivial and thus there is more than one solution. This is consistent with the adjoint approach.

Example 6.3. Consider the operator $\mathscr{A}$ defined in Example 5.7. In particular, $\mathscr{A} : \mathbb{U} = \text{Span}\{1,\ x,\ x^2\} \subset \mathbb{U} = \mathbb{L}^2(-1, 1) \to \mathbb{R}^2$ such that

$$u(x) \in \mathbb{U} \mapsto \mathscr{A}u = \begin{bmatrix} u(x_1) \\ u(x_2) \end{bmatrix} \in \mathbb{R}^2,$$

where x_1 and x_2 are two given points in $(-1, 1)$. We wish to know the existence and uniqueness of a solution of the equation $\mathscr{A}u = \boldsymbol{f}$ for a given $\boldsymbol{f} \in \mathbb{R}^2$. Note that this is an operator setting for the problem of fitting a quadratic polynomial in $u(x)$ with two pieces of information about $u(x)$. Based on elementary knowledge, we know that there is more than one quadratic as we can only hope to determine a quadratic uniquely with three pieces of information. We expect the proposed adjoint approach to provide the same answer. From Example 5.7 the adjoint was found to be

$$\mathscr{A}^* = [\delta(x - x_1), \delta(x - x_2)],$$

$$\delta(x - y) = 15\frac{(3y^2 - 1)}{8}x^2 + 3\frac{y}{2}x + 3\frac{(3 - 5y^2)}{8}.$$

It follows that

$$\begin{aligned} \mathscr{A}^*\boldsymbol{v} = 15&\frac{\left[3\left(x_1^2\boldsymbol{v}_1 + x_2^2\boldsymbol{v}_2\right) - (\boldsymbol{v}_1 + \boldsymbol{v}_2)\right]}{8}x^2 \\ &+ 3\frac{(x_1\boldsymbol{v}_1 + x_2\boldsymbol{v}_2)}{2}x + 3\frac{\left[3(\boldsymbol{v}_1 + \boldsymbol{v}_2) - 5\left(x_1^2\boldsymbol{v}_1 + x_2^2\boldsymbol{v}_2\right)\right]}{8}. \end{aligned} \tag{6.2}$$

For the existence, we look at the null space of $\mathscr{A}^*$, that is, all $\boldsymbol{v}$ such that $\mathscr{A}^*\boldsymbol{v} = 0$ for all x. This gives us three equations for $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, and clearly the only solution is $\boldsymbol{v}_1 = \boldsymbol{v}_2 = 0$ when $x_1 \neq x_2$. In this case, there is always a solution for any $\boldsymbol{f} \in \mathbb{R}^2$ as $\mathsf{R}(\mathscr{A}) = [\mathsf{N}(\mathscr{A}^*)]^\perp = \{\boldsymbol{0}\}^\perp = \mathbb{R}^2$. On the other hand, if $x_1 = x_2$, a general solution is $\boldsymbol{v}_1 = -\boldsymbol{v}_2$. In this case,

$$\mathsf{N}(\mathscr{A}^*) = span\left\{[1, -1]^T\right\},$$

and thus $\mathscr{A}u = \boldsymbol{f}$ has a solution iff $\boldsymbol{f} \in \mathsf{R}(\mathscr{A}) = \mathsf{N}(\mathscr{A}^*)^{\perp} = span\left\{[1,1]^T\right\}$.

For the uniqueness, we look at the range space of $\mathscr{A}^*$. From (6.2), given x_1 and x_2, the specific forms of coefficients of the second order polynomial $\mathscr{A}^*\boldsymbol{v}$ can range at most a subset of $\mathbb{U}$ when $\boldsymbol{v}$ varies in $\mathbb{R}^2$. Thus $\mathsf{R}(\mathscr{A}^*) \subset \mathbb{U}$ and this implies that $\mathsf{N}(\mathscr{A})$, the orthogonal complement of $\mathsf{R}(\mathscr{A}^*)$ in $\mathbb{U}$, is nontrivial. We conclude that there are many solutions. This is consistent with the direct approach based on $\mathsf{N}(\mathscr{A})$ as fixing a quadratic polynomial $u(x)$ at two values $u(x_1)$ and $u(x_2)$ does not uniquely identity a polynomial.

We would like to point out that whether the adjoint approach is preferable is problem-specific. Sometimes it is trivial to see the answers directly from $\mathscr{A}$ instead of $\mathscr{A}^*$.

Example 6.4. We continue Example 5.8 here where $\mathscr{A} : \mathbb{U} := \mathbb{P}^n[0,1] \subset \mathbb{L}^2(0,1) \to \mathbb{U}$ as

$$\mathscr{A}u := xu' := x\frac{du}{dx}.$$

The adjoint was found to be

$$\mathscr{A}^*v = \delta(x-1)\,v(1) - (xv)',$$

where $\delta(x-1)$ is an nth-order polynomial in $\mathbb{U}$ defined via

$$\int_0^1 u(x)\overline{\delta(x-1)}\,dx = u(1), \quad \forall u \in \mathbb{U}.$$

We are interested in knowing if there is a solution to the equation $\mathscr{A}u = f$ for a given $f \in \mathbb{U}$.

For the existence of a solution, we have a choice to look at $\mathsf{R}(\mathscr{A})$ or $\mathsf{N}(\mathscr{A}^*)$. While it is not trivial to characterize $\mathsf{N}(\mathscr{A}^*)$, it is much easier to know what $\mathsf{R}(\mathscr{A})$ looks like. Indeed, from the definition of $\mathscr{A}$, $\mathscr{A}u$ is clearly an nth-order polynomial without the zero-order term. As a result, $\mathscr{A}u = f$ has a solution iff f is any polynomial of order at most n without zero-order term. In other words, $\mathscr{A}u = f$ has no solution if f is a constant.

Now suppose f is such that $\mathscr{A}u = f$ has a solution, the question is if it is unique. Since it is not clear if $\mathsf{R}(\mathscr{A}^*)$ is a proper subset of $\mathbb{U}$, we look at $\mathsf{N}(\mathscr{A})$. Let $u \in \mathsf{N}(\mathscr{A})$, we have

$$x\frac{du}{dx} = 0, \quad \forall x \in [0,1],$$

which implies that all coefficients in the polynomial u vanish except the constant term. In other words, any constant u resides in $\mathsf{N}(\mathscr{A})$. We conclude that there is more than one solution (just simply add a constant to a solution we obtain another solution) since $\mathsf{N}(\mathscr{A})$ is not trivial.

Theorem 6.1. *Let* $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ *be a linear operator and* $\dim(\mathbb{U}), \dim(\mathbb{V}) < \infty$.

Chapter 7
Eigenvalue problem for self-adjoint operators

Abstract In this chapter, we consider linear operators mapping a space into itself, i.e. $\mathscr{A} : \mathbb{X} \to \mathbb{X}$. We say that $\mathbb{U} \subseteq \mathbb{X}$ is an invariant subspace of $\mathscr{A}$ if $\mathscr{A}(\mathbb{U}) \subseteq \mathbb{U}$. Trivial examples for invariant subspaces are $\mathbb{U} = \{\theta\}$, $\mathbb{U} = \mathsf{N}(\mathscr{A})$, and $\mathbb{U} \equiv \mathbb{X}$. The question is if there are other (non-trivial) invariant subspaces of $\mathscr{A}$? It turns out the eigensubspaces, when exist, are invariant subspaces. The existence of eigensubspaces is guaranteed when $\mathscr{A}$ self-adjoint, i.e. $\mathscr{A}^* = \mathscr{A}$. In that case, any invariant subspace of $\mathscr{A}$, including the trivial ones except $\{\theta\}$, is a combination of eigensubspaces. In other words, eigensubspaces determine all invariant subspaces of $\mathscr{A}$. Furthermore, we shall see that a self-adjoint linear operator is completely characterized by its eigenpairs. Discuss the application of eigenvalue problems in

- model reduction
- vibration analysis (Jeff's work)
- orthogonal polynomial, Fourier series, Sturm-Liouville problems
- SVD
- stability of ODEs (eigenvalues must have non-positive real parts), and iterative methods (eigenvalues with magnitude less than one)

7.1 Eigenvalue problem for self-adjoint operators

Definition 7.1 (Eigenvalue problem). Let $\mathscr{A} : \mathbb{X} \to \mathbb{X}$ be a linear operator. The problem of finding a pair $\{\lambda, u\}$ such that

$$\mathscr{A}u = \lambda u \qquad \text{where } \lambda \in \mathbb{C},\ u \in \mathbb{X} \tag{7.1}$$

is called an eigenvalue problem if there exists a nontrivial pair $\{\lambda, u\}$ (x *is not a zero vector/function but* λ *could be zero*). In particular:

- λ is called an eigenvalue.

- u is called an eigenfunction, associated with the eigenvalue λ, of $\mathscr{A}$. If $\mathbb{X}$ is a finite-dimensional space, i.e. $\dim \mathbb{X} < \infty$, u is typically called an eigenvector. We shall use eigenfunction and eigenvector interchangeably.

Definition 7.2 (Self-adjoint operator). If $\mathscr{A}^* = \mathscr{A}$, then $\mathscr{A}$ is called self-adjoint.

Lemma 7.1. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{X}$ be a linear operator and $\mathscr{A}$ is self-adjoint. Then:*

1. *Eigenvalues of $\mathscr{A}$ are real.*
2. *Eigenfunctions corresponding to distinct eigenvalues are orthogonal to each other. That is, if $\{\lambda, u\}$ and $\{\alpha, v\}$ are two eigen-pairs and $\lambda \neq \alpha$ then $(u, v)_{\mathbb{X}} = 0$.*

Proof. For the first assertion, we have $\lambda (u, v)_{\mathbb{X}} = (\lambda u, u)_{\mathbb{X}} = (\mathscr{A} u, u)_{\mathbb{X}} = (u, \mathscr{A} u)_{\mathbb{X}} = (u, \lambda u)_{\mathbb{X}} = \overline{\lambda} (u, u)_{\mathbb{X}}$. Thus, $\left(\lambda - \overline{\lambda}\right) (u, \lambda u)_{\mathbb{X}} = 0$, and this implies $\lambda = \overline{\lambda}$, or λ is real. For the second assertion, we observe that $\lambda (u, v)_{\mathbb{X}} = (\mathscr{A} u, v)_{\mathbb{X}} = (u, \mathscr{A} v)_{\mathbb{X}} = \alpha (u, v)_{\mathbb{X}}$. Therefore, $(\lambda - \alpha) (u, v)_{\mathbb{X}} = 0$, and this implies $(u, v)_{\mathbb{X}} = 0$.

Remark 7.1. Note that Definition 7.1, Definition 7.2, and Lemma 7.1 hold for both finite and infinite dimensional cases.

Proposition 7.1. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{X}$ be a linear operator. Then, $\mathscr{A}$ has at least one eigenvalue.*

Proof. Let n be the dimension of $\mathbb{X}$ and $\boldsymbol{E}$ be a basis. Let $\boldsymbol{A}$ and $\boldsymbol{u}$ be the matrix and vector presentation of $\mathscr{A}$ and u in the basis $\boldsymbol{E}$. The matrix representation of the eigenvalue problem(7.1) reads

$$\boldsymbol{A}\boldsymbol{u} = \lambda \boldsymbol{u},$$

that is, λ is an eigenvalue of $\mathscr{A}$ iff it is also an eigenvalue of $\boldsymbol{A}$. Since $\det (\boldsymbol{A} - \lambda I) = 0$ has n roots for λ (including repeated ones), there is at least one eigenvalue.

7.2 Spectral decomposition of self-adjoint operators in finite-dimensional spaces

Theorem 7.1. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{X}$ be a self-adjoint linear operator. Then, an orthonormal basis of $\mathbb{X}$ can be constructed from eigenfunctions of $\mathscr{A}$.*

Proof. Proposition 7.1 implies that $\mathscr{A}$ has at least one eigenfunctions. Let $\mathcal{S}$ be the span of all eigenfunctions of $\mathscr{A}$. Owing to Lemma 7.1, it is sufficient to

show that $\mathcal{S} = \mathbb{X}$. If $\mathcal{S}^\perp = \{\theta\}$, then clearly $\mathcal{S} = \mathbb{X}$. Now suppose $\mathcal{S}^\perp \neq \{\theta\}$. If $u \in \mathcal{S}^\perp$ and $\{\lambda, v\}$ be an eigen-pair of $\mathscr{A}$, then $(\mathscr{A}u, v)_{\mathbb{X}} = (u, \mathscr{A}v)_{\mathbb{X}} = \lambda\,(u, v)_{\mathbb{X}} = 0$ since $v \in \mathcal{S}$. Thus $\mathscr{A} : \mathcal{S}^\perp \to \mathcal{S}^\perp$. By Proposition 7.1, $\mathscr{A}$ has an eigenfunction w in $\mathcal{S}^\perp$, which means that $w \in \mathcal{S} \cap \mathcal{S}^\perp$, which in turn implies $w = \theta$, a contradiction. It follows that $\mathcal{S}^\perp = \{\theta\}$ and this concludes the proof.

Corollary 7.1 (Spectral decomposition of self-adjoint operators in finite dimensions). *Let* $\dim \mathbb{X} = n$ *and* $\mathscr{A} : \mathbb{X} \to \mathbb{X}$ *be a linear and self-adjoint operator. There exists* n *real values,* $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_n$ *and orthonormal vectors* $u_1, u_2, \ldots u_n$ *such that:*

1. $\mathscr{A}u_i = \lambda_i u_i$.
2. *For any* $x \in \mathbb{X}$ *we have*

$$\mathscr{A}x = \sum_{i=1}^{n} \lambda_i\,(x, u_i)_{\mathbb{X}}\,u_i, \implies \mathscr{A} = \sum_{i=1}^{n} \lambda_i\,(\cdot, u_i)_X\;\; u_i,$$

that is, $\mathscr{A}$ *is completely characterized by its eigenpairs.*

Proof. The first assertion is clear due to Lemma 7.1, Proposition 7.1, and Theorem 7.1. The second assertion is also obvious due to (thanks to Theorem 7.1)

$$x = \sum_{i=1}^{n} (x, u)_{\mathbb{X}}\,u_i, \tag{7.2}$$

the linearity of $\mathscr{A}$, and the definition of eigenpairs.

Corollary 7.2. *Consider the same setting in Corollary 7.1. Let* $\boldsymbol{E}$ *be an orthonormal basis of* $\mathbb{X}$. *Let* $\boldsymbol{A}$ *and* $\boldsymbol{u}_i$ *be the matrix and vector representations of* $\mathscr{A}$ *and* $u_i, i = 1, \ldots, n$. *Then,* $\{\lambda_i, u_i\}_{i=1}^n$ *are eigenpairs of* $\mathscr{A}$ *iff* $\{\lambda_i, \boldsymbol{u}_i\}_{i=1}^n$ *are eigenpairs of* $\boldsymbol{A}$.

Example 7.1 (Eigen-decomposition of self-adjoint (Hermitian) matrices). Let $\boldsymbol{A} : \mathbb{F}^n \to \mathbb{F}^n$ be a self-adjoint matrix. Applying either Corollary 7.1 or Corollary 7.2 we conclude that $\boldsymbol{A}$ has n real eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \cdots \geq \lambda_n$ and orthonormal eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_n$ such that $\boldsymbol{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ and

$$\boldsymbol{A} = \sum_{i=1}^{n} \lambda_i\,(\cdot, \boldsymbol{u}_i)_{\mathbb{F}^n}\,\boldsymbol{u}_i = \sum_{i=1}^{n} \lambda_i \boldsymbol{u}_i \boldsymbol{u}_i^* = \boldsymbol{U}\Lambda\boldsymbol{U}^*,$$

where $\boldsymbol{U} \in \mathbb{F}^{n\times n}$ is the eigenmatrix whose columns are eigenvectors of $\boldsymbol{A}$ and Λ is a diagonal matrix whose diagonals are the corresponding eigenvalues of $\boldsymbol{A}$. Thus, the standard eigendecomposition for self-adjoint matrices is a special case of Corollary 7.1.

Example 7.2. Consider Example 5.8 again but with $n = 2$. We are interested in finding the eigenpairs for $\mathscr{A}$. To that end, we deploy Corollary 7.2 to

first find the eigenpairs of the matrix representation of $\mathscr{A}$ in an orthonormal basis of $\mathbb{P}^1[0,1]$, and then form the eigenpairs for $\mathscr{A}$. An orthogonal basis for $\mathbb{P}^2[0,1]$ is clearly $\{1, 2x-1\}$. The matrix representation $\boldsymbol{A}$ of $\mathscr{A}$ in this basis is given as

$$\boldsymbol{A} = \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

for which the eigenpairs are $\left\{0, \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$ and $\left\{1, \begin{bmatrix} \sqrt{2}/2 \\ \sqrt{2}/2 \end{bmatrix}\right\}$. Thus, the eigenpairs for $\mathscr{A}$ are $\{0, 1\}$ and $\left\{1, \sqrt{2}x\right\}$.

7.3 Orthogonal projection and self-adjointness

We next discuss the relationship between self-adjointness, orthogonal projection, and eigenspaces of a self-adjoint operator.

Definition 7.3 (Orthogonal projection). A linear operator $\mathscr{P} : \mathbb{X} \to \mathbb{X}$ is a projection if $\mathscr{P}^2 := \mathscr{P}\mathscr{P} = \mathscr{P}$. If, in addition, $\mathsf{R}(\mathscr{P}) \perp \mathsf{N}(\mathscr{P})$, then $\mathscr{P}$ is an orthogonal projection.

Proposition 7.2. *Let $\mathbb{X}$ be an inner product space. A projection $\mathscr{P} : \mathbb{X} \to \mathbb{X}$ is orthogonal iff $\mathscr{P}$ is self-adjoint.*

Proof. For any $x \in \mathsf{N}(\mathscr{P})$ we have

$$(\mathscr{P}y, x)_{\mathbb{X}} = (y, \mathscr{P}^* x)_{\mathbb{X}} = (y, \mathscr{P}x)_{\mathbb{X}} = 0, \quad \forall y \in \mathbb{X},$$

which ends the proof. Another way to see this is to use the result $\mathsf{N}(\mathscr{P}) \perp \mathsf{R}(\mathscr{P}^*)$ from the closed range Theorem 5.2, but we omit the details for brevity.

It is easy to verify that any orthogonal projection $\mathscr{P}$ orthogonally projects $\mathbb{X}$ into $\mathsf{R}(\mathscr{P})$ while

$$\mathscr{I} - \mathscr{P},$$

where $\mathscr{I}$ is the identity operator, is the orthogonal projection into $\mathsf{N}(\mathscr{P})$. Indeed, by the self-adjointness, we have

$$(\mathscr{P}x, y - \mathscr{P}y)_{\mathbb{X}} = \left(x, \mathscr{P}y - \mathscr{P}^2 y\right)_{\mathbb{X}} = 0, \quad \forall x, y \in \mathbb{X}.$$

We thus have the following generalized Pythagorean theorem for any $x \in \mathbb{X}$ and any orthogonal projection $\mathscr{P}$:

$$\|x\|_{\mathbb{X}}^2 = \|(\mathscr{I} - \mathscr{P})\, x\|_{\mathbb{X}}^2 + \|\mathscr{P}x\|_{\mathbb{X}}^2 . \tag{7.3}$$

We will see in Theorem 11.1 that among all vectors in $\mathsf{R}(\mathscr{P})$, $\mathscr{P}x$ is the closest one to x.

Corollary 7.3. *Let $\mathscr{P} : \mathbb{X} \to \mathbb{X}$ be an orthogonal projection into an inner product space $\mathbb{X}$. The following holds*

- *$\mathscr{P}$ is a continuous linear map with $\|\mathscr{P}\| = 1$.*
- *1 and 0 are the only eigenvalues of $\mathscr{P}$. The former corresponds to the projection of $\mathsf{R}(\mathscr{P})$ onto itself, and the latter to the projection of $\mathsf{N}(\mathscr{P})$ to $\{\theta\}$.*

Proof. The first assertion is a direct consequence of the Pythagorean identity (7.3). For the second assertion, we let (λ, x) be an eigenpair of $\mathscr{P}$. We have

$$\lambda x = \mathscr{P}x = \mathscr{P}\mathscr{P}x = \lambda \mathscr{P}x = \lambda^2 x,$$

which concludes the proof.

The following are some standard results of an orthogonal projection in an inner product space.

Lemma 7.2. *Let $\mathscr{P} : \mathbb{X} \to \mathbb{X}$ be an orthogonal projection into an inner product space $\mathbb{X}$. Show that*

- *$\mathsf{R}(\mathscr{P})$ and $\mathsf{N}(\mathscr{P})$ are closed linear subspaces of $\mathbb{X}$.*
- *$\mathsf{R}(\mathscr{P}) = \mathsf{N}(\mathscr{P})^{\perp}$ and $\mathsf{N}(\mathscr{P}) = \mathsf{R}(\mathscr{P})^{\perp}$.*
- *For any $x \in \mathbb{X}$, there is a unique $r \in \mathsf{R}(\mathscr{P})$ and a unique $n \in \mathsf{N}(\mathscr{P})$ such that $x = r + n$ and furthermore $\|x\|_{\mathbb{X}}^2 = \|r\|_{\mathbb{X}}^2 + \|n\|_{\mathbb{X}}^2$. That is, $\mathsf{R}(\mathscr{P}) \oplus \mathsf{N}(\mathscr{P}) = \mathbb{X}$.*

Proof. For the first assertion, the linear space structure of $\mathsf{R}(\mathscr{P})$ and $\mathsf{N}(\mathscr{P})$ is clear by the linearity of $\mathscr{P}$. The closeness is due to the continuity of $\mathscr{P}$ (see Corollary 7.3). The third assertion is the direct consequence of the second and the Pythagorean identity (7.3). *We prove the second assertion as it exposes the self-adjointness of an orthogonal projection.* Take an arbitrary $x \in \mathsf{N}(\mathscr{P})$. Since any $y \in \mathsf{R}(\mathscr{P})$ can be written as $y = \mathscr{P}z$ for some $z \in \mathbb{X}$, we have

$$(x, y) = (x, \mathscr{P}z) = (\mathscr{P}^* x, z) = (\mathscr{P}x, z) = 0,$$

which means $\mathsf{N}(\mathscr{P}) \subseteq \mathsf{R}(\mathscr{P})^{\perp}$. On the other hand, take $x \in \mathsf{R}(\mathscr{P})^{\perp}$, we have

$$0 = (x, \mathscr{P}y) = (\mathscr{P}^* x, y) = (\mathscr{P}x, y), \quad \forall y \in \mathbb{X},$$

which gives $\mathscr{P}x = \theta$, and thus by definition $x \in \mathsf{N}(\mathscr{P})$, that is, $\mathsf{R}(\mathscr{P})^{\perp} \subseteq \mathsf{N}(\mathscr{P})$, which concludes the proof.

Let us also record an important result on the unique corresponding between an orthogonal projection and a closed linear subspace of a Hilbert space (see, e.g., [109, Theorem 5.16.4]).

Theorem 7.2. *Let $\mathcal{S}$ be a closed linear subspace of a Hilbert space $\mathbb{X}$. There exists a unique orthogonal projection $\mathscr{P} : \mathbb{X} \to \mathcal{S}$ such that $\mathsf{R}(\mathscr{P}) = \mathcal{S}$.*

We would like to point out that up to this point, the results of this section is valid in both finite and infinite dimensional Hilbert spaces.

Inspired by the spectral decomposition of a self-adjoint operator in Corollary 7.1, we define

$$\mathscr{P} := \sum_{i=1}^{n} (\cdot, u_i)_{\mathbb{X}}\, u_i, \tag{7.4}$$

where $\{u_1, \cdots, u_n\}$ is an orthonormal basis of $\mathbb{X}$. It is easy to verify that: i) $\mathscr{P}^2 := \mathscr{P}\mathscr{P} = \mathscr{P}$, and ii) $\mathscr{P}$ is self-adjoint. Thus $\mathscr{P}$ is an orthogonal projection onto $\mathbb{X}$: in fact, $\mathscr{P}$ is the identity operator in this case, and (7.4) *is known as a resolution of identity for n-dimensional spaces*. More generally, we can easily see (see Problem 7.1) that

$$\mathscr{P}^r := \sum_{i=1}^{r} (\cdot, u_i)_{\mathbb{X}}\, u_i, \tag{7.5}$$

where $r \leq n$, is an orthogonal projection into $\mathsf{R}(\mathscr{P}^r)$ spanned by $\{u_1, \cdots, u_r\}$.

Remark 7.2. Thus, in order to compute an orthogonal projection onto a subspace of a Hilbert space, we can first find an orthonormal set and form a projection operator similar to (7.5). We will generalize this result to infinite dimensional settings in Corollary 16.1.

Problems

7.1. We continue with the setting in Problem 5.3 but now let $\mathbb{Y} = \mathbb{X}$ and $\{\lambda, u\}$ be an eigenpair of $\mathscr{A}$. Suppose that $\mathscr{A}$ is a bijection. Show that $\{1/\lambda, u\}$ is an eigenpair of $\mathscr{A}^{-1}$.

7.2. Prove Corollary 7.2.

7.3. Provide a detailed solution of Example 7.2.

7.4. Let $\mathscr{P} : X \to X$ be a linear bounded projection. Show that $\mathscr{P}$ is orthogonal iff it is normal, i.e., $\mathscr{P}\mathscr{P}^* = \mathscr{P}^*\mathscr{P}$.

Problem 7.1. Let $\{u_1, \cdots, u_r\}$, $r \in \mathbb{N}$, be an orthonormal set in a Hilbert space $\mathbb{X}$. Define.

$$\mathscr{P}^r := \sum_{i=1}^{r} (\cdot, u_i)_{\mathbb{X}}\, u_i,$$

where $r \leq n$. Show that $\mathscr{P}^r$ is an orthogonal projection onto $\mathsf{R}(\mathscr{P}^r)$ spanned by $\{u_1, \cdots, u_r\}$. If $\mathbb{X}$ is n-dimensional Hilbert space, then show that $\mathscr{P}^n$ is a resolution of identity.

Chapter 8
The singular value decomposition (SVD) from adjoint perspective

Abstract The singular value decomposition (SVD) is perhaps the most popular factorization of a linear operator.[1] Its popularity lies on its indispensable role in theories and applications of mathematics. An early history of SVD can be found in [134] where the author discussed its birth from linear algebra to functional analysis by five mathematicians: Beltrami (1835-1899), Jordan (1838-1921), Sylvester (1814-1897), Schmidt (1876-1959), and Weyl (1885-1955). Some [140] believed that the proof for rectangular matrices was due to Eckart and Young [49]. A short summary of extraodinary wide range of applications can be found in [100] where we can find applications of SVD ranging from crystal growth to data compression and to politics. The extension of SVD from two dimensional tensors (such as matrices) to higher dimensional tensors is discussed in details in [85]. [77] [123] Intuitive interpretation of SVD vectors Should include the reference to two books that we have here at home. The prerequisites for this chapter are:

- Linear operator (basic definitions including range and null spaces: see Chapter 5)
- Basics on Hilbert spaces (inner product, orthogonality)
- Linear algebra
- Chapter Chapter 7

Consider $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ with $\dim(\mathbb{X}) = n$ and $\dim(\mathbb{Y}) = m$. It is clear that the operators $\mathscr{A}^*\mathscr{A} : \mathbb{X} \to \mathbb{X}$ and $\mathscr{A}\mathscr{A}^* : \mathbb{Y} \to \mathbb{Y}$ are linear and self-adjoint. The spectral decomposition of self-adjoint operator in Corollary 7.1 states that $\exists \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$, $\lambda_i \in \mathbb{R}$, $i = 1, \ldots, n$, and an orthonormal basis $u_1,\ u_2, \ldots, u_n$ of $\mathbb{X}$ such that

$$\mathscr{A}^*\mathscr{A}u_i = \lambda_i u_i,$$

which implies, after taking the inner product both sides with u_i,

[1] A Google search at the time of writing returns 29,300,000 results: way far beyond any factorizations/decompositions.

$$\|\mathscr{A}u_i\|_{\mathbb{Y}}^2 = \lambda_i \|u_i\|_{\mathbb{Y}}^2,$$

and thus $\lambda_i \geq 0$. We define $\sigma_i^2 = \lambda_i$ and have

$$\mathscr{A}^*\mathscr{A}u_i = \sigma_i^2 u_i. \tag{8.1}$$

Theorem 8.1 (Singular Value Decomposition (SVD)). *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ with $\dim(\mathbb{X}) = n$ and $\dim(\mathbb{Y}) = m$. Then, there exist $\{\sigma_i, u_i, v_i\}$ (the singular triplets of $\mathscr{A}$) with $\sigma_1 \geq \ldots \geq \sigma_i \geq \cdots \geq \sigma_k \geq 0$ and $k = min\{n, m\}$, an orthonormal basis $\{u_1, u_2, \ldots, u_n\}$ of $\mathbb{X}$, and an orthonormal basis $\{v_1, v_2, \ldots, v_m\}$ of $\mathbb{Y}$ such that:*

1. *$\mathscr{A}u_i = \sigma_i v_i$ for $i = 1, \ldots, r$ and $\mathscr{A}u_i = \theta$ for $i = r+1, \ldots, n$,*
2. *$\mathscr{A}^* v_j = \sigma_j u_j$ for $j = 1, \ldots, r$ and and $\mathscr{A}^* v_j = \theta$ for $j = r+1, \ldots, m$,*
3. *$\mathscr{A}$ is completely determined by its singular triplets in the following sense: for any $x \in \mathbb{X}$, we have*

$$\mathscr{A}x = \sum_{i=1}^{r} \sigma_i \left(x, u_i\right)_{\mathbb{X}} v_i, \implies \mathscr{A}\left(\cdot\right) = \sum_{i=1}^{r} \sigma_i \left(\cdot, u_i\right)_{\mathbb{X}} v_i,$$

where r is the maximum index for which $\sigma_r > 0$.

Proof. Starting from (8.1), let r be the maximum index for which $\sigma_r > 0$ and define $v_i = \dfrac{1}{\sigma_i}\mathscr{A}u_i$ for $i \leq r$, so that

$$\mathscr{A}u_i = \sigma_i v_i. \tag{8.2}$$

Substituting (8.2) into (8.1) gives

$$\mathscr{A}^* v_i = \sigma_i u_i. \tag{8.3}$$

We claim that $\left\{\sigma_i^2, v_i\right\}$ for $i \leq r$ are the eigenpairs of the self-adjoint operator $\mathscr{A}\mathscr{A}^*$. To see this, applying $\mathscr{A}$ to both sides of (8.3) to arrive at

$$\mathscr{A}\mathscr{A}^* v_i = \sigma_i \mathscr{A}u_i = \sigma_i^2 v_i. \tag{8.4}$$

That is, for every eigenpair of $\mathscr{A}^*\mathscr{A}$ corresponding to a non-zero eigenvalue we have an eigenpair of $\mathscr{A}\mathscr{A}^*$ with the same eigenvalue. By the same token, we can show that for every eigenpair of $\mathscr{A}\mathscr{A}^*$ corresponding to a non-zero eigenvalue we have an eigenpair of $\mathscr{A}^*\mathscr{A}$ with the same eigenvalue. As a result, the rest of eigenvalues of $\mathscr{A}^*\mathscr{A}$ and $\mathscr{A}\mathscr{A}^*$ with indices larger than r must be 0. The orthonormality of $\{u_1, u_2, \ldots, u_n\}$ and $\{v_1, v_2, \ldots, v_m\}$ is the direct consequence of the spectral decomposition of self-adjoint operators in Corollary 7.1. The third assertion is clear owing to the first assertion and (7.2).

It follows that the SVD of $\mathscr{A}^*$ is given as (see Problem 8.2)

$$\mathscr{A}^*(\cdot) = \sum_{i=1}^{r} \sigma_i (\cdot, v_i)_{\mathbb{Y}} u_i.$$

Corollary 8.1. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ with $\dim(\mathbb{X}) = n$ and $\dim(\mathbb{Y}) = m$ and its singular triplets given in Theorem 8.1. Show that*

$$\|\mathscr{A}\| := \sup_{x \in \mathbb{X}} \frac{\|\mathscr{A}x\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}} = \|\mathscr{A}^*\| := \sup_{y \in \mathbb{Y}} \frac{\|\mathscr{A}^* y\|_{\mathbb{X}}}{\|y\|_{\mathbb{Y}}} = \sigma_1,$$

and $x = u_1$ and $y = v_1$ achieve the suprema.

Proof. From Theorem 8.1 we have

$$\|\mathscr{A}x\|_{\mathbb{X}}^2 = \sum_{i=1}^{r} \sigma_i^2 |(x, u_i)_{\mathbb{X}}|^2 \leq \sigma_1^2 \sum_{i=1}^{r} |(x, u_i)_{\mathbb{X}}|^2 \leq \sigma_1^2 \sum_{i=1}^{n} |(x, u_i)_{\mathbb{X}}|^2 = \sigma_1^2 \|x\|_{\mathbb{X}}^2.$$

The proof concludes by noticing that the equality happens when $x = u_1$. The proof for $\mathscr{A}^*$ is similar by using the SVD of $\mathscr{A}^*$.

8.1 The application of SVD for the closed range theorem, the rank-nullity theorem, and the fundamental theorem of linear algebra

The SVD decomposition in Theorem 8.1 allows us to provide trivial proofs of various important results in finite dimensions including the closed range Theorem 5.2, the rank-nullity theorem and the fundamental theorem of linear algebra. While these results are typically presented for matrices, it is not more difficult to do so for generic linear operators using our general setting as we shall show. To begin, we note that $u_{r+1}, \ldots, u_n$ are orthonormal eigenfunctions corresponding to 0 eigenvalues of A^*A. We conclude that

$$span\{u_{r+1}, \ldots, u_n\} = \mathsf{N}(\mathscr{A}^*\mathscr{A}) = \mathsf{N}(\mathscr{A}),$$

which, together with the fact that $span\{u_1, \ldots, u_n\} = \mathbb{X}$ implies

$$span\{\mathscr{A}u_1, \ldots, \mathscr{A}u_r\} = \mathsf{R}(\mathscr{A}),$$

which in turn yields

$$span\{v_1, \ldots, v_r\} = \mathsf{R}(\mathscr{A}),$$

since $\mathscr{A}u_i = \sigma_i v_i$, $i = 1, \ldots, r$.

Similarly, we have

$$span\{v_{r+1}, \ldots, v_m\} = \mathsf{N}(\mathscr{A}\mathscr{A}^*) = \mathsf{N}(\mathscr{A}^*),$$

and

$$span\{u_1, \ldots, u_r\} = \mathsf{R}(\mathscr{A}^*).$$

With these conclusions in hand, the assertions in the closed range Theorem 5.2 and Corollary 5.2 are now trivial.

The SVD also provides an obvious proof for the rank-nullity theorem [13] as

$$\underbrace{\dim(\mathbb{X})}_{=n} = \underbrace{\dim(\mathsf{N}(\mathscr{A}))}_{=n-r} + \underbrace{\dim(\mathsf{R}(\mathscr{A}))}_{=r}, \tag{8.5a}$$

and

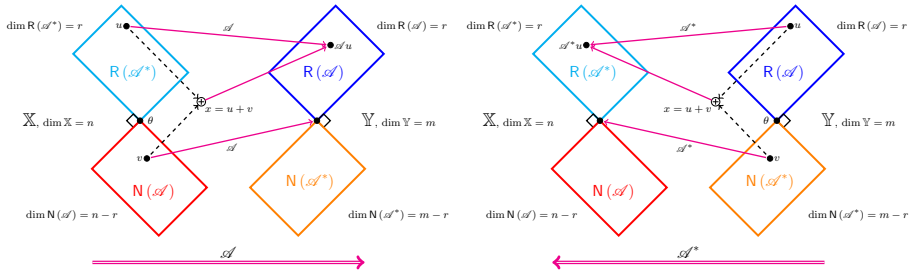
$$\underbrace{\dim(\mathbb{Y})}_{=m} = \underbrace{\dim(\mathsf{N}(\mathscr{A}^*))}_{=m-r} + \underbrace{\dim(\mathsf{R}(\mathscr{A}^*))}_{=r}, \tag{8.5b}$$

which, together with the closed range Theorem 5.2, is the basis for the fundamental theorem of linear algebra advocated by Strang [136]. This is demonstrated in Figure 8.1 which shows the important role of the four fundamental subspaces $\mathsf{R}(\mathscr{A}), \mathsf{N}(\mathscr{A}), \mathsf{R}(\mathscr{A}^*)$, and $\mathsf{N}(\mathscr{A}^*)$ on the operation of $\mathscr{A}$ and its adjoint $\mathscr{A}^*$. In particular, $\mathscr{A}$ only acts on the range space of $\mathscr{A}^*$, and the results of its action, the range space of $\mathscr{A}$, coincides with the orthogonal complement of the nullspace of $\mathscr{A}^*$. Conversely, $\mathscr{A}^*$ only acts on the range space of $\mathscr{A}$ and the results of its action, the range space of $\mathscr{A}^*$, coincides with the orthogonal complement of the nullspace of $\mathscr{A}$. In other words, the characterization of $\mathscr{A}^*$ completely determines the action of $\mathscr{A}$ and vice versa. We will see the intertwine of $\mathscr{A}$ and $\mathscr{A}^*$ again in Chapter 15 from a different perspective. We emphasize that if we remove the dimensions and assume that $\mathsf{R}(\mathscr{A})$ is closed, then, thanks to the closed range theorem Theorem 5.2, the two diagrams in Figure 8.1 also hold for infinite dimensional Hilbert spaces.

Clearly, at the heart of the SVD is the eigenvalue decomposition (8.4), which could be challenging if it is analytically not tractable on the original operators. In that case, one has to resort to numerical methods. For finite-dimensional settings, an easier path is to explore the matrix representation of the linear operator.

Corollary 8.2 (SVD through matrix representation). *Consider* $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ *with* $\dim(\mathbb{X}) = n$ *and* $\dim(\mathbb{Y}) = m$, *and* $\boldsymbol{E}$ *and* $\boldsymbol{G}$ *be orthonormal bases of* $\mathbb{X}$ *and* $\mathbb{Y}$, *respectively. Let* $\{\sigma_i, u_i, v_i\}$ *be the singular triplets of* $\mathscr{A}$ *with* $1 \leq i \leq k = min\{n,\ m\}$ *where* $\{u_1, u_2, \ldots, u_n\}$ *and* $\{v_1, v_2, \ldots, v_m\}$ *be orthonormal bases of* $\mathbb{X}$ *and* $\mathbb{Y}$, *respectively, given in Theorem 8.1. Denote* $\boldsymbol{u}$ *and* $\boldsymbol{v}$ *as the coordinate vectors of* u *and* v *in the bases* $\boldsymbol{E}$ *and* $\boldsymbol{G}$, *respectively, and* $\boldsymbol{A}$ *as the matrix representation of* $\mathscr{A}$ *with respect to the bases* $\boldsymbol{E}$ *and* $\boldsymbol{G}$. *Then* $\{\sigma_i, \boldsymbol{u}_i, \boldsymbol{v}_i\}$ *be the singular triplets of* $\boldsymbol{A}$ *with*

1. $\boldsymbol{A}\boldsymbol{u}_i = \sigma_i \boldsymbol{v}_i$ *for* $i = 1, \ldots, r$ *and* $\boldsymbol{A}\boldsymbol{u}_i = \theta$ *for* $i = r+1, \ldots, n$,



(a) Operation of $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$. (b) Operation of $\mathscr{A}^* : \mathbb{Y} \to \mathbb{X}$.

Fig. 8.1: The fundamental theorem of algebra: four fundamental subspaces $\mathsf{R}(\mathscr{A}), \mathsf{N}(\mathscr{A}), \mathsf{R}(\mathscr{A}^*), \mathsf{N}(\mathscr{A}^*)$, and the operation of $\mathscr{A}$ and $\mathscr{A}^*$ viewed from these subspaces. If we remove the dimensions and assume that $\mathsf{R}(\mathscr{A})$ is closed, then the two diagrams also hold for infinite dimensional Hilbert spaces.

2. *$\boldsymbol{A}^* \boldsymbol{v}_j = \sigma_j \boldsymbol{u}_j$ for $j = 1, \ldots, r$ and and $\boldsymbol{A}^* \boldsymbol{v}_j = \theta$ for $j = r+1, \ldots, m$, where $\boldsymbol{A}^*$ is the conjugate tranpose of $\boldsymbol{A}$.*
3. *$\boldsymbol{A}$ is completely determined by its singular triplets in the following sense: for any $\boldsymbol{x} \in \mathbb{F}^n$, we have*

$$\boldsymbol{A}\boldsymbol{x} = \sum_{i=1}^{r} \sigma_i \boldsymbol{u}_i^* \boldsymbol{x}\, \boldsymbol{v}_i, \implies \boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{v}_i \boldsymbol{u}_i^*,$$

where r is the maximum index for which $\sigma_r > 0$.

Conversely, if $\{\sigma_i, \boldsymbol{u}_i, \boldsymbol{v}_i\}$, with $1 \le i \le k = min\{n,\ m\}$, are the singular triplets of $\boldsymbol{A}$, then (σ_i, u_i, v_i) are the singular triplets of $\mathscr{A}$.

Proof. The result is obvious owing to the matrix representation of linear operator and the coordinate vector of a vector in the corresponding bases (see Proposition 5.4), and the fact that two vectors are orthonormal iff their coordinate vectors in an orthogonal basis are orthonormal (as, e.g., $(\boldsymbol{u}_i, \boldsymbol{u}_j)_{\mathbb{F}^n} = (u_i, v_j)_{\mathbb{X}}$).

Example 8.1 (SVD of matrices). For a matrix $\boldsymbol{A} : \mathbb{X} = \mathbb{R}^n \to \mathbb{Y} = \mathbb{R}^m$ and if we choose $\boldsymbol{E}$ and $\boldsymbol{G}$ as the canonical bases for $\mathbb{R}^n$ and $\mathbb{R}^m$ with the standard Euclidean inner products, respectively, then the matrix representation of $\boldsymbol{A}$ is itself and thus the SVD of $\boldsymbol{A}$ is given by Corollary 8.2. In this case, $\boldsymbol{A}^* = \boldsymbol{A}^T$. Furthermore: i) $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$ are orthonormal eigenvectors of $\boldsymbol{A}^T\boldsymbol{A}$; ii) $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$ are orthonormal eigenvectors of $\boldsymbol{A}\boldsymbol{A}^T$; iii) $\sigma_i^2, \quad i = 1, \ldots r$, are nonzero eigenvalues of $\boldsymbol{A}^T\boldsymbol{A}$ or $\boldsymbol{A}\boldsymbol{A}^T$; and iv) from $\boldsymbol{A} = \sum_{i=1}^{r} \sigma_i \boldsymbol{v}_i \boldsymbol{u}_i^*$ we can write the full SVD form

$$\boldsymbol{A} = \underbrace{\begin{bmatrix} | & | & & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_r & \dots & \boldsymbol{v}_m \\ | & | & & | & & | \end{bmatrix}}_{\boldsymbol{V}} \underbrace{\begin{bmatrix} \sigma_1 & & & & & \\ & \sigma_2 & & & & \\ & & \ddots & & & \\ & & & \sigma_r & & \\ & & & & 0 & \\ & & & & & \ddots \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} | & | & & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_r & \dots & \boldsymbol{u}_n \\ | & | & & | & & | \end{bmatrix}^T}_{\boldsymbol{U}^T},$$

that is,

$$\boldsymbol{A} = \boldsymbol{V}\Sigma\boldsymbol{U}^T,$$

or the reduced SVD form

$$\boldsymbol{A} = \underbrace{\begin{bmatrix} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \dots & \boldsymbol{v}_r \\ | & | & & | \end{bmatrix}}_{\boldsymbol{V}_r} \underbrace{\begin{bmatrix} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{bmatrix}}_{\Sigma_r} \underbrace{\begin{bmatrix} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \dots & \boldsymbol{u}_r \\ | & | & & | \end{bmatrix}^T}_{\boldsymbol{U}_r^T},$$

that is,

$$\boldsymbol{A} = \boldsymbol{V}_r\Sigma_r\boldsymbol{U}_r^T.$$

Example 8.2. Now consider the operator $\mathscr{A} : \mathbb{U} = \text{Span}\{1, x, x^2\} \subset \mathbb{X} = \mathbb{L}^2(-1,1) \to \mathbb{R}^2$ defined in Example 5.6. We are going to find the singular value decomposition of $\mathscr{A}$ indirectly via its matrix representation using Corollary 8.2. Clearly, two orthonormal bases for $\mathbb{U}$ and $\mathbb{R}^2$ are $\boldsymbol{E} = \left\{1, x, \frac{1}{2}\left(3x^2-1\right)\right\}$, $\boldsymbol{G} = \left\{[1,0]^T, [0,1]^T\right\}$, respectively. It is a simple exercise to show that the matrix representation $\boldsymbol{A}$ of $\mathscr{A}$ in these two bases is given by

$$\boldsymbol{A} = \begin{bmatrix} 2 & 0 & 1 \\ 2 & 4/3 & 1/3 \end{bmatrix}$$

and from Proposition 5.4 we know that the matrix presentation $\boldsymbol{A}^*$ of the adjoint $\mathscr{A}^*$ is $\boldsymbol{A}^* = \boldsymbol{A}^T$. From the proof of Theorem 8.1, by computing the eigendecomposition of $\boldsymbol{A}\boldsymbol{A}^T$ and $\boldsymbol{A}^T\boldsymbol{A}$, we can find the full SVD of $\boldsymbol{A}$ as

$$\boldsymbol{A} = \underbrace{\begin{bmatrix} -0.6701 & 0.7423 \\ 0.7423 & -0.6701 \end{bmatrix}}_{\boldsymbol{V}} \underbrace{\begin{bmatrix} 3.1306 & 0 & 0 \\ 0 & 1.0433 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} -0.9023 & 0.1385 & -0.4082 \\ -0.3162 & -0.8564 & 0.4082 \\ -0.2931 & 0.4974 & 0.8165 \end{bmatrix}^T}_{\boldsymbol{U}^T}.$$

Now, Corollary 8.2 shows that the singular values of $\mathscr{A}$ are $\{3.1306, 1.0433\}$ together with the left and right singular functions $u_i = \boldsymbol{u}_i(1) + \boldsymbol{u}_i(2)x + \frac{\boldsymbol{u}_i(3)}{2}\left(3x^2-1\right)$, $i = 1,2,3$, and $\boldsymbol{v}_j$, $j = 1,2$, respectively.

8.2 From SVD to the principle component analysis and the proper orthogonal decomposition

What we are going to show is that using SVD provides a simple derivation of the principle component analysis (PCA) [59, 74] (a.k.a the proper orthogonal decomposition, POD, [16, 42, 81]). The principle component analysis is a linear dimensional reduction in which we are given an operator[2] $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$, where $\dim \mathbb{X} = n$ and $\dim \mathbb{Y} = m$ and the task is the find a vector (*the principle component*) $\phi \in \mathbb{X}$ along which $\mathscr{A}$ magnify the most. That is, we would like to find ϕ such that

$$\max_{\phi \in \mathbb{X}} \frac{\|\mathscr{A}\phi\|_{\mathbb{Y}}}{\|\phi\|_{\mathbb{X}}}, \tag{8.6}$$

which, as can be seen from (5.6), is nothing more than the operator norm (or spectral norm) of the operator $\mathscr{A}$. We can write (8.6) equivalently as

$$\max_{\phi \in \mathbb{X}} \frac{(\mathscr{A}\phi, \mathscr{A}\phi)_{\mathbb{Y}}}{(\phi, \phi)_{\mathbb{X}}}. \tag{8.7}$$

Note that in the POD approach [16], $\mathscr{A}\phi$ is the action of projecting ϕ on data ensemble members and $(\mathscr{A}\phi, \mathscr{A}\phi)_{\mathbb{Y}}$ is an assemble squared average of ϕ, up to a constant. Thus, (8.7) is exactly the POD task in which we look for a unit vector $\phi/\sqrt{(\phi, \phi)_{\mathbb{X}}}$ in $\mathbb{X}$ (the data space) such that it aligns most with the data ensemble on average. Thus, the principle vector is the one that aligns most with the data.

In other words, we have shown that

$$u_1 = \arg\max_{\phi \in \mathbb{X}} \frac{\|\mathscr{A}\phi\|_{\mathbb{Y}}}{\|\phi\|_{\mathbb{X}}},$$

and

$$\max_{\phi \in \mathbb{X}} \frac{\|\mathscr{A}\phi\|_{\mathbb{Y}}}{\|\phi\|_{\mathbb{X}}} = \sigma_1.$$

Note that rigorously we must write

$$u_1 \in \arg\max_{\phi \in \mathbb{X}} \frac{\|\mathscr{A}\phi\|_{\mathbb{Y}}}{\|\phi\|_{\mathbb{X}}}$$

[2] In the literature, $\mathscr{A}$ is typically a matrix (or assemble) of data, but our exposition allows us to work directly with a general linear operator in finite-dimensional spaces without incurring additional difficulties. Furthermore, $\mathscr{A}$ is not required to be a mapping between finite-dimensional spaces for the results of this section. In fact, all we need is that $\mathscr{A}$ is a compact operator (see Theorem 14.2).

since it could be that $\sigma_1 = \sigma_2$. In that case, either u_1 or u_2 (or u_3 if $\sigma_1 = \sigma_2 = \sigma_3$, etc) is a solution. We then can simply take u_1. Thus, we can ignore this little technicality without loss of generality.

The task of finding the next principle component can be written as

$$\max_{\substack{\phi \in \mathbb{X} \\ (\phi, u_1)_{\mathbb{X}} = 0}} \frac{\|\mathscr{A}\phi\|_{\mathbb{Y}}}{\|\phi\|_{\mathbb{X}}},$$

which, by following the same derivation as above, is equivalent to

$$\max_{\substack{\|\boldsymbol{\alpha}\|_{\mathbb{C}^n} = 1 \\ (\boldsymbol{\alpha}, \boldsymbol{e}_1)_{\mathbb{C}^n} = 0}} \sum_{j=1}^{r} \sigma_j^2 \left|\boldsymbol{\alpha}_j\right|^2.$$

Clearly, the condition $(\boldsymbol{\alpha}, \boldsymbol{e}_1)_{\mathbb{C}^n} = 0$ implies that $\boldsymbol{\alpha} = [0, \boldsymbol{\beta}]^T$, where $\boldsymbol{\beta} \in \mathbb{C}^{n-1}$. Thus, the problem of finding the second principle component now reads

$$\max_{\|\boldsymbol{\beta}\|_{\mathbb{C}^{n-1}} = 1} \sum_{j=2}^{r} \sigma_j^2 \left|\boldsymbol{\beta}_{j-1}\right|^2,$$

and as similarly shown above, the optimal solution is $\boldsymbol{\beta} = [1, 0, \ldots, 0]^T$ and thus the second principle component is $\phi = u_2$ and

$$\max_{\substack{\phi \in \mathbb{X} \\ (\phi, u_1)_{\mathbb{X}} = 0}} \frac{\|\mathscr{A}\phi\|_{\mathbb{Y}}}{\|\phi\|_{\mathbb{X}}} = \sigma^2.$$

By induction, we see that the left singular vectors $u_1, u_2, \ldots, u_r$, in this order, are the most to the least principle components of the operator $\mathscr{A}$. Carrying out the same procedure for the adjoint operator $\mathscr{A}^*$ we conclude that the right singular vectors $v_1, v_2, \ldots, v_r$ are the most to the least principle components for $\mathscr{A}^*$. The above exposition also explains why SVD, POD, and PCA are equivalent.

8.3 Application of SVD in pseudo-inverse

Recall that we are interested in linear map $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$, where $\dim \mathbb{X} = n$ and $\dim \mathbb{Y} = m$. We consider the cases where $\mathscr{A}$ is not bijective and thus not invertible and seek to define a more general inverse called pseudo-inverse [106, 115, 44].

Definition 8.1 (Pseudo-inverse). Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$. $\mathscr{A}^\dagger : \mathbb{Y} \to \mathbb{X}$ is called a pseudo-inverse of $\mathscr{A}$ if it satisfies

ps1) $\mathscr{A}^\dagger \mathscr{A} x = x$ for all $x \in \mathsf{R}(\mathscr{A}^*) = \mathsf{N}(\mathscr{A})^\perp$,

ps2) $\mathscr{A}^\dagger y = \theta$ for all $y \in \mathsf{N}(\mathscr{A}^*) = \mathsf{R}(\mathscr{A})^\perp$, and
ps3) If $y_1 \in \mathsf{R}(\mathscr{A})$ and $y_2 \in \mathsf{N}(\mathscr{A}^*)$, then $\mathscr{A}^\dagger(y_1 + y_2) = \mathscr{A}^\dagger y_1 + \mathscr{A}^\dagger y_2$.

Remark 8.1. Note that due to the Corollary 5.2 we have

$$\mathsf{R}(\mathscr{A}) = \mathscr{A}\left(\mathsf{N}(\mathscr{A})^\perp \oplus \mathsf{N}(\mathscr{A})\right) = \mathscr{A}\left(\mathsf{N}(\mathscr{A})^\perp\right),$$

and thus Item **ps1)** defines $\mathscr{A}^\dagger$ on $\mathsf{R}(\mathscr{A})$ and as the left-inverse of $\mathscr{A}$ when restricted to $\mathsf{N}(\mathscr{A})^\perp$. The second Item **ps2)** defines the nullspace of $\mathscr{A}^\dagger$ as the nullspace of $\mathscr{A}^*$ due to $\mathbb{Y} = \mathsf{N}(\mathscr{A}^*) \oplus \mathsf{R}(\mathscr{A})$ from Corollary 5.2. The third Item **ps3)** ensures that $\mathscr{A}^\dagger$ is a linear operator on $\mathbb{Y}$ as a direct consequence of Corollary 5.2 and Item **ps1)** together with Item **ps2)**. The uniqueness of $\mathscr{A}^\dagger$ can be seen as follows. For any $y \in \mathbb{Y}$, Corollary 5.2 tells us that there are unique $x_1 \in \mathsf{R}(\mathscr{A}^*)$ and $y_2 \in \mathsf{N}(\mathscr{A}^*)$ such that $y = \mathscr{A}x_1 + y_2$ and we have

$$\mathscr{A}^\dagger y = \mathscr{A}^\dagger \mathscr{A} x_1 = x_1.$$

As a consequence, $\mathscr{A}^\dagger$ is unique as if there were another operator $\mathscr{B}$ satisfying all the above conditions, we would have

$$\left(\mathscr{A}^\dagger - \mathscr{B}\right) y = 0, \quad \forall y \in \mathbb{Y},$$

and hence $\mathscr{A}^\dagger = \mathscr{B}$.

We would like to point out that the pseudo-inverse Definition 8.1 and Remark 8.1 are also valid for any closed range bounded linear operator $\mathscr{A}$ between two arbitrary Hilbert spaces $\mathbb{X}$ and $\mathbb{Y}$ [44].

Now let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, where $r \leq \min\{n, m\}$, $\{u_i\}_{i=1}^n$, and $\{v_i\}_{i=1}^m$ be the the singular triplets of $\mathscr{A}$ from Theorem 8.1, and recall

$$\mathscr{A}x = \sum_{i=1}^{r} \sigma_i (x, u_i)_{\mathbb{X}} v_i, \quad \forall x \in \mathbb{X}.$$

From section 8.1 we know that $\mathsf{N}(\mathscr{A})^\perp = span\{u_1, \cdots, u_r\}$, and thus for any $x \in \mathsf{N}(\mathscr{A})^\perp$ we have $x = \sum_{i=1}^{r} (x, u_i) u_i$. From definition of $\mathscr{A}^\dagger$, we must have

$$\mathscr{A}^\dagger \mathscr{A} x = \sum_{i=1}^{r} (x, u_i)_{\mathbb{X}} \mathscr{A}^\dagger (\sigma_i v_i) = \sum_{i=1}^{r} (x, u_i)_{\mathbb{X}} u_i = x, \quad \forall x \in \mathsf{N}(\mathscr{A})^\perp,$$

which implies

$$\mathscr{A}^\dagger v_i = \frac{1}{\sigma_i} u_i, \quad i = 1, \ldots, r.$$

Since section 8.1 also tells us that $\mathsf{R}(\mathscr{A}) = span\{v_1, \ldots, v_r\}$, and any $y \in \mathsf{R}(\mathscr{A})$ is expressed as

$$y = \sum_{i=1}^{r} (y, v_i)_{\mathbb{Y}} \, v_i.$$

Thus, $\forall y \in \mathsf{R}(\mathscr{A})$ we have

$$\mathscr{A}^{\dagger} y = \sum_{i=1}^{r} (y, v_i)_{\mathbb{Y}} \, \mathscr{A}^{\dagger} v_i = \sum_{i=1}^{r} \frac{1}{\sigma_i} (y, v_i)_{\mathbb{Y}} \; u_i,$$

which is also valid for all $y \in \mathbb{Y}$ since $\mathbb{Y} = \mathsf{N}(\mathscr{A}^*) \oplus \mathsf{R}(\mathscr{A})$ and $\mathsf{N}\left(\mathscr{A}^{\dagger}\right) = \mathsf{N}(\mathscr{A}^*)$ as discussed above. This implies

$$\mathscr{A}^{\dagger}(\cdot) = \sum_{i=1}^{r} \frac{1}{\sigma_i} (\cdot, v_i)_{\mathbb{Y}} \; u_i,$$

which is exactly the SVD of $\mathscr{A}^{\dagger}$. In other words, we have shown that $\sigma_r^{-1} \geq \sigma_{r-1}^{-1} \geq \ldots \geq \sigma_1^{-1} > 0$, where $r \leq \min\{n, m\}$, $\{v_i\}_{i=1}^{m}$, and $\{u_i\}_{i=1}^{n}$ are the singular triplets of $\mathscr{A}^{\dagger}$. By construction, $\mathscr{A}^{\dagger}$ satisfies the three conditions in Definition 8.1, and thus is the unique pseudo-inverse that we are looking for. Let us formally summarize the results Theorem 8.1, the SVD of $\mathscr{A}^*$ in (8.9), and the above derivation for the SVD of $\mathscr{A}^{\dagger}$.

Lemma 8.1. *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, where $r \leq \min\{n, m\}$, $\{u_i\}_{i=1}^{n}$, and $\{v_i\}_{i=1}^{m}$ be the the singular triplets of $\mathscr{A}$, that is,*

$$\mathscr{A}(\cdot) = \sum_{i=1}^{r} \sigma_i (\cdot, u_i)_{\mathbb{X}} \, v_i, . \tag{8.8}$$

then the SVD of $\mathscr{A}^*$ is given as

$$\mathscr{A}^*(\cdot) = \sum_{i=1}^{r} \sigma_i (\cdot, v_i)_{\mathbb{Y}} \, u_i, \tag{8.9}$$

and the SVD of $\mathscr{A}^{\dagger}$ is given as

$$\mathscr{A}^{\dagger}(\cdot) = \sum_{i=1}^{r} \frac{1}{\sigma_i} (\cdot, v_i)_{\mathbb{Y}} \; u_i. \tag{8.10}$$

Lemma 8.1 immediately implies a few important results (properties 2, 3, and 4 are also valid for re also valid for any closed range bounded linear operator $\mathscr{A}$):

1. $(\mathscr{A}^*)^* = \mathscr{A}$,

2. $\left(\mathscr{A}^\dagger\right)^\dagger = \mathscr{A}$,
3. $\left(\mathscr{A}^\dagger\right)^* = (\mathscr{A}^*)^\dagger$,
4. $(\alpha\mathscr{A})^\dagger = \alpha^{-1}\mathscr{A}^\dagger, \quad \forall \alpha \neq 0$, and
5. if we know the SVD for either one of the operators, we know the SVD of the others.

As we have shown the SVD of a linear operator not only allows us to constructively derive the pseudo-inverse of a linear operator but also provides easy proofs for most results regarding the pseudo-inverse, some of which are presented below.

Lemma 8.2. *The following hold:*

1. $\mathscr{A}\mathscr{A}^\dagger\mathscr{A} = \mathscr{A}$,
2. $\mathscr{A}^\dagger\mathscr{A}\mathscr{A}^\dagger = \mathscr{A}^\dagger$,
3. $\left(\mathscr{A}^\dagger\mathscr{A}\right)^* = \mathscr{A}^\dagger\mathscr{A}$, *and* $\left(\mathscr{A}\mathscr{A}^\dagger\right)^* = \mathscr{A}\mathscr{A}^\dagger$,
4. $\mathscr{A}\mathscr{A}^\dagger$ *is the orthogonal projection onto* $\mathsf{R}(\mathscr{A})$. *Similarly,* $\mathscr{A}^\dagger\mathscr{A}$ *is the orthogonal projection onto* $\mathsf{R}(\mathscr{A}^*)$,
5. $\mathscr{A}^\dagger = \mathscr{A}^*\left(\mathscr{A}\mathscr{A}^*\right)^{-1}$ *if* $\dim\left(\mathsf{R}\left(\mathscr{A}^*\right)\right) = m$,
6. $\mathscr{A}^\dagger = \left(\mathscr{A}^*\mathscr{A}\right)^{-1}\mathscr{A}^*$ *if* $\dim\left(\mathsf{R}\left(\mathscr{A}\right)\right) = n$.

Proof. With the SVD of $\mathscr{A}$, $\mathscr{A}^*$, and $\mathscr{A}^\dagger$ in Lemma 8.1, the proof of these assertions is straightforward and we show a couple of them here. For the first assertion, we have

$$\begin{aligned}\mathscr{A}\mathscr{A}^\dagger\mathscr{A}x = \sum_{i=1}^{r}\sigma_i\left(x,u_i\right)_{\mathbb{X}}\mathscr{A}\mathscr{A}^\dagger v_i &= \sum_{i=1}^{r}\sigma_i\left(x,u_i\right)_{\mathbb{X}}\mathscr{A}\left(\sigma_i^{-1}u_i\right)\\ &= \sum_{i=1}^{r}\sigma_i\left(x,u_i\right)_{\mathbb{X}}v_i = \mathscr{A}x, \quad \forall x\in\mathbb{X},\end{aligned}$$

and thus the first assertion holds.

For the third assertion, we have, for any $x \in \mathbb{X}$:

$$\mathscr{A}^\dagger\mathscr{A}x = \sum_{i=1}^{r}\frac{1}{\sigma_i}\left(\sum_{j=1}^{r}\sigma_j\left(x,u_j\right)_{\mathbb{X}}v_j, v_i\right)_{\mathbb{Y}}u_i = \sum_{j=1}^{r}\left(x,u_j\right)_{\mathbb{X}}u_j, \tag{8.11}$$

but

$$\mathscr{A}^*\left(\mathscr{A}^\dagger\right)^* x = \sum_{j=1}^{r}\sigma_j\left(\sum_{i=1}^{r}\frac{1}{\sigma_i}\left(x,u_i\right)_{\mathbb{X}}v_i, v_j\right)_{\mathbb{Y}}u_j = \sum_{j=1}^{r}\left(x,u_j\right)_{\mathbb{X}}u_j,$$

and thus the first part of the third assertion holds. The second part follows similarly.

The third assertion says that $\mathscr{A}^\dagger\mathscr{A}$ is self-adjoint. On the other hand,

$$\left(\mathscr{A}^{\dagger}\mathscr{A}\right)^{2} = \mathscr{A}^{\dagger}\mathscr{A}\mathscr{A}^{\dagger}\mathscr{A} = \mathscr{A}^{\dagger}\mathscr{A},$$

where we have used the first assertion in the last inequality. By Proposition 7.2 and (8.11), we conclude that $\mathscr{A}\mathscr{A}^{\dagger}$ is an orthogonal projection onto $span\{u_1, \ldots, u_r\} = \mathsf{R}\{\mathscr{A}^*\}$ (see section 8.1) and this proves the fourth assertion.

For the fifth assertion, we assume that $\dim(\mathsf{R}(\mathscr{A}^*)) = m$. In this case, from section 8.1 we know that

$$\mathsf{R}(\mathscr{A}^*) = span\{u_1, \ldots, u_m\},$$

and thus $r = m$ in (8.9) with $\sigma_m > 0$. This, together with Theorem 8.1, implies that $n \geq m$ and $r = m$ in both (8.9) and (8.8). We thus have

$$\mathscr{A}\mathscr{A}^* y = \sum_{i=1}^{m} \sigma_i \left(\sum_{j=1}^{m} \sigma_j \left(y, v_j\right) u_j, u_i \right) v_i = \sum_{j=1}^{m} \sigma_j^2 \left(y, v_j\right) v_j,$$

which implies that $\mathscr{A}\mathscr{A}^* : \mathbb{Y} \to \mathbb{Y}$ is bijective, and thus invertible (see Problem 8.5). The fifth assertion is now equivalent to

$$\mathscr{A}^{\dagger}\mathscr{A}\mathscr{A}^* = \mathscr{A}^*,$$

but this is trivially true from Item **ps1)** since

$$\mathscr{A}^{\dagger}\mathscr{A} \underbrace{\mathscr{A}^* y}_{\text{a member of } \mathsf{R}(\mathscr{A}^*)} = \mathscr{A}^* y, \quad \forall y \in \mathbb{Y}.$$

We would like to point out that properties 1, 2, 3, and 4 of Lemma 8.2 are also valid for any closed range bounded linear operator $\mathscr{A}$.

Problems

Problem 8.1. Consider $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$. Show that $\mathscr{A}^*\mathscr{A}$ and $\mathscr{A}\mathscr{A}^*$ are self-adjoint.

8.1. Explain why the proof of closed range Theorem 5.2 for finite dimensional setting is trivial using SVD.

Problem 8.2. Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, where $r \leq \min\{n, m\}$, $\{u_i\}_{i=1}^n$, and $\{v_i\}_{i=1}^m$ be the the singular triplets of $\mathscr{A}$. Show that $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, where $r \leq \min\{n, m\}$, $\{v_i\}_{i=1}^m$, and $\{u_i\}_{i=1}^n$ are the the singular triplets of $\mathscr{A}^*$, and the SVD of $\mathscr{A}^*$ is given in (8.9).

Problem 8.3. Let $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ where $\mathbb{U}$ and $\mathbb{V}$ are finite dimensions. Show that

$$\|\mathscr{A}\| := \sup_{u \in \mathbb{X}} \frac{\|\mathscr{A}u\|_{\mathbb{Y}}}{\|u\|_{\mathbb{X}}} = \sigma_1,$$

where σ_1 is the largest singular value of $\mathscr{A}$.

Problem 8.4. Show that the left singular vectors of an operator $\mathscr{A} : X \to \mathbb{Y}$, where $\mathbb{X}$ and $\mathbb{Y}$ are finite dimensional Hilbert spaces, are the principle components of $\mathscr{A}^*$.

Problem 8.5. Let $\mathscr{A} : X \to \mathbb{Y}$ with $\dim \mathbb{X} = n$ and $\dim \mathbb{Y} = m$. Show that if $\dim(\mathsf{R}(\mathscr{A}^*)) = n$, then $\mathscr{A}\mathscr{A}^* : \mathbb{Y} \to \mathbb{Y}$ is invertible.

Problem 8.6. Derive the adjoint $\mathbb{R}^*$ in section 22.3

Problem 8.7 (The method of snapshots). Let us consider $\boldsymbol{A} \in \mathbb{R}^{m \times n}$. A practical problem of interest is to find the dominant, or principle, directions (those associated with largest singular) of the column space and where $m \gg n$. Finding the SVD decomposition of a large $\boldsymbol{A}$ could be computationally challenging. The idea of the method of snapshots [133] is to find it indirectly via the eigendecomposition of a much smaller matrix. The first step is to compute the eigenvalue decomposition of $\boldsymbol{A}^T\boldsymbol{A}$ so that we have r eigenvectors $\{\boldsymbol{u}_i\}_{i=1}^r$ corresponding to the largest eigenvalues $\{\sigma_i\}_{i=1}^r$. Then form the first r dominant directions as $\boldsymbol{v}_i = \dfrac{\boldsymbol{A}\boldsymbol{u}_i}{\sigma_i}$, $i = 1, \ldots, r$. Show that $\{\boldsymbol{v}_i\}_{i=1}^r$ are indeed the orthonormal principle directions of the column space of $\boldsymbol{A}$. We would like to point out that these principle directions are also known as the principle components in the principle component analysis (PCA) [59, 74]. They are also widely known as the proper orthogonal decomposition (POD) modes [17] or the Karhunen-Loeve (KL) modes [42, 81].

Chapter 9
Efficient constrained optimization with adjoint

Abstract The field of optimization is vast (see, e.g., [97, 19, 111, 18] and the references therein) and we restrict ourselves to unconstrained optimization problems and constrained optimization problems with equality constraints. We start the chapter with a brief development of the first and second-order necessary conditions for optimality of unconstrained optimization problems in one dimension in section 9.1. The beauty here is that the extension to any dimension is simply a corollary. The rest of the chapter focuses on the first-order necessary conditions. This is one of the chapters in which we develop abstract theory (valid for both finite and infinite dimensions) even though we are in a finite-dimensional part. The reason is that separating them can only invite unnecessary repetitions. The expense is that we have to cope with some (hopefully not) advanced concepts including dual spaces and the Fréchet derivatives, but the payoff is worthwhile as we shall have a single set of results that is valid for all dimensions. In particular, section 9.2 develops the first-order optimality condition for an abstract unconstrained optimization problem. In section 9.3, we first heuristically develop the first-order optimality condition for optimization problem with equality constraints, and then rigorously derive the first-order optimality conditions based on inverse function theorem. The proof for an abstract Lagrangian multiplier theorem—the foundation of the adjoint method—becomes straightforward with an application of the closed range Theorem 5.2. We end the chapter with section 9.4 on an important class of separable optimization problems with equality constraints in which we derive the reduced-space approach using adjoint.

9.1 Unconstrained optimization in one dimension

We begin our development with unconstrained optimization in finite-dimensional spaces. Let $f : \mathbb{R}^n \ni \boldsymbol{u} \mapsto f(\boldsymbol{u}) \in \mathbb{R}$ and we are interested in studying the optimization problem $\min_{\boldsymbol{u}\in\mathbb{R}^n} f(\boldsymbol{u})$. It is sufficient to consider the case $n = 1$ as

results for optimization problems in higher dimensions (including infinite dimensions, at least the first order optimality conditions) follow as corollaries. We focus on local optimization problems.

Definition 9.1. v is a (local) minimizer of $f(u)$ if there exists a open neighborhood, i.e. $\boldsymbol{B}_\delta(v) := \{w : |w - v| < \delta\}$ (ball with radius δ centered at v in $\mathbb{R}$), for some $\delta > 0$, such that

$$f(u) \geq f(v), \quad \forall u \in \boldsymbol{B}_\delta(v).$$

The problem at hand is to find the necessary and sufficient conditions for v to be a minimizer. To that end, we consider the Taylor remainder theorem [10, 84] which states that for twice-differentiable function $f(u)$ and for any $\varepsilon \in \mathbb{R}$, there exists $0 < \theta < 1$ such that

$$f(v+\varepsilon) = f(v) + \varepsilon \underbrace{f'(v)}_{\text{gradient } g(v)} + \frac{1}{2}\varepsilon^2 \underbrace{f''(v+\theta\varepsilon)}_{\text{Hessian } h(v+\theta\varepsilon)}. \tag{9.1}$$

If v is a minimizer, what can we say about the gradient $g(\cdot)$ and the Hessian $h(\cdot)$ at v? We are interested in only the necessary conditions. Here is an answer.

Lemma 9.1 (First and second order necessary conditions for optimality in $\mathbb{R}$). *Suppose $f(u) : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable in a neighborhood of a minimizer v. It is necessary that*

i) the gradient vanishes, i.e., $g(v) = 0$, and
ii) the Hessian is non-negative, i.e., $h(v) \geq 0$.

Proof. We carry out the proof by contradiction. For the first assertion, we suppose that $f'(v) < 0$ and note that we can pick[1] $\varepsilon > 0$ such that

$$\varepsilon f'(v) + \frac{\varepsilon^2}{2} f''(v+\theta\varepsilon) < 0,$$

and together with (9.1) we conclude that $f(v+\varepsilon) < f(v)$: a contradiction. A similar contradiction argument can be carry out if $f'(v) > 0$. Thus $f'(v) = 0$.

For the second assertion, suppose $h(v) = f''(v) < 0$. By continuity of $f''(u)$ we can choose sufficiently small $|\varepsilon|$ such that $f''(v+\theta\varepsilon) < 0$. Then (9.1) reduces to

$$f(y+\varepsilon) = f(v) + \frac{1}{2}\varepsilon^2 f''(y+\theta\varepsilon) < f(v),$$

[1] Due to the continuity of $f''(u)$, we can define $M := \max_{u\in[v-L,v+L]} |f''(u)|$, for some sufficiently large $L > 0$, and then simply pick some $0 < \varepsilon < -2\frac{f'(v)}{M}$.

which is a contradiction, and this concludes the proof.

Corollary 9.1. *Suppose $f(\boldsymbol{u}) : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable in a neighborhood of a minimizer $\boldsymbol{v}$, where $\mathbb{R}^n$ is equipped with the standard Euclidean inner product. It is necessary that*

i) the gradient vanishes, i.e., $\dfrac{\partial f}{\partial \boldsymbol{u}_i}(\boldsymbol{v}) = 0$, $i = 1, \ldots, n$ and
ii) the Hessian matrix is semi-positive definite, i.e., $\boldsymbol{H}(\boldsymbol{v}) \geq 0$.

Proof. We prove the first assertion (the first order optimality condition) as the second one follows similarly. Note that argument is general and will be used again in Lemma 9.2 to derive the first order optimality condition in general vector spaces.

$$
\begin{aligned}
&\boldsymbol{v} \text{ is a minimizer of } f(\boldsymbol{u}) \\
&\Downarrow && \text{by definition} \\
&f(\boldsymbol{u}) \geq f(\boldsymbol{v}), \quad \forall \boldsymbol{u} \in \boldsymbol{B}_\delta(\boldsymbol{v}) := \{\boldsymbol{w} \in \mathbb{R}^n : \|\boldsymbol{w} - \boldsymbol{v}\|_{\mathbb{R}^n} < \delta\} \\
&\Downarrow && \text{pick an arbitrary } \boldsymbol{v} \\
&F(\varepsilon) := f(\boldsymbol{v} + \varepsilon \boldsymbol{v}) \geq f(\boldsymbol{v}), \quad \forall \varepsilon \in \boldsymbol{B}_{\delta/\|\boldsymbol{v}\|_{\mathbb{R}^n}}(0) := \{\kappa : |\kappa| < \delta/\|\boldsymbol{v}\|_{\mathbb{R}^n}\} \\
&\Downarrow && \text{by definition} \\
&0 \text{ is a minimizer of } F(\varepsilon) \\
&\Downarrow && \text{by Lemma 9.1} \\
&\left.\frac{dF}{d\varepsilon}\right|_{\varepsilon=0} = 0 \\
&\Downarrow && \text{by chain rule} \\
&\sum_{i=1}^{n} \frac{\partial f}{\partial \boldsymbol{u}_i}(\boldsymbol{v})\, \boldsymbol{v}_i = 0 \\
&\Downarrow && \boldsymbol{v} \text{ is arbitrary} \\
&\frac{\partial f}{\partial \boldsymbol{u}_i}(\boldsymbol{v}) = 0, \quad i = 1, \ldots, n.
\end{aligned}
$$

9.2 First-order optimality condition for unconstrained optimizations in any dimensions

Since our goal is to *establish the necessary conditions for optimality that is valid for both finite and infinite dimensional settings*, we present a systematic approach on abstract vector space to accomplish this. **To the end of this section, unless otherwise stated, the results are valid for both finite and infinite dimensional settings**. We begin with the notion of the dual space $\mathbb{U}^*$ consisting of linear and bounded functionals on $\mathbb{U}$ (see Definition 5.5). For $\ell \in \mathbb{U}^*$ and $u \in \mathbb{U}$, we use the standard duality pairing

$$\langle \ell, u \rangle_{\mathbb{U}^* \times \mathbb{U}} \equiv \ell(u)$$

to denote the action of ℓ on u (or the evaluation of ℓ at u): see Remark 5.4. *For simplicity in writing, we shall conventionally use* $\langle \ell, u \rangle_{\mathbb{U}}$ *to denote a duality pairing instead of* $\langle \ell, u \rangle_{\mathbb{U}^* \times \mathbb{U}}$. The object of interest is nonlinear function on a vector space $\mathbb{U}$, i.e. *functional*:

$$f : \mathbb{U} \ni u \mapsto f(u) \in \mathbb{R}.$$

The classical derivatives are not well-defined in this case, and this asks for an extension of derivatives in vector spaces. Though there are other extensions in the literature (such as Gâteaux derivative), let us focus on the Fréchet derivative extension (see, e.g., [97, 11]), which relies on bounded linear maps (see Chapter 5).

Definition 9.2. Suppose that there is a linear and bounded map $\mathscr{D}f(u;\cdot) : \mathbb{U} \to \mathbb{R}$ such that

$$f(u+v) = f(u) + \mathscr{D}f(u;v) + o\left(\|v\|_{\mathbb{U}}\right), \tag{9.2}$$

where the standard little-oh notation means

$$\lim_{\|v\|_{\mathbb{U}} \to 0} \frac{o\left(\|v\|_{\mathbb{U}}\right)}{\|v\|_{\mathbb{U}}} = 0.$$

Then $\mathscr{D}f(u;\cdot)$ is called the Fréchet derivative of the functional $f(\cdot)$ at u, and *we say* $f(\cdot)$ *is Fréchet differentiable* at u.

When the Fréchet derivative exists, we can compute it conveniently as

$$\mathscr{D}f(u;v) = \left.\frac{df}{dt}(u+tv)\right|_{t=0} = \lim_{t \to 0} \frac{f(u+tv) - f(u)}{t}$$

For convenience, we use $\mathscr{D}f(u)$ to denote the Fréchet derivative $\mathscr{D}f(u;\cdot)$ when the argument is irrelevant. It is important to note that by definition the Fréchet derivative $\mathscr{D}f(u)$ resides in $\mathbb{U}^*$ and thus we interchangeably write it in the duality pairing form

$$\mathscr{D}f(u;v) = \langle \mathscr{D}f(u), v \rangle_{\mathbb{U}}.$$

Due to the linear nature of $\mathscr{D}f(u)$, we also write

$$\mathscr{D}f(u)\, v := \langle \mathscr{D}f(u), v \rangle_{\mathbb{U}}.$$

The following is a generalization of the elementary introduction to Fréchet derivative in Example 5.4.

Definition 9.3 (Fréchet gradient). Let $f : \mathbb{U} \to \mathbb{R}$. The gradient of $f(\cdot)$ at u, denoted as $\nabla f(u) \in \mathbb{U}$, is defined as a function on $\mathbb{U}$ such that

$$(\nabla f(u), v)_{\mathbb{U}} = \mathscr{D}f(u;v) = \langle \mathscr{D}f(u), v \rangle_{\mathbb{U}} = \mathscr{D}f(u)\, v, \quad \forall v \in \mathbb{U}.$$

That is, we define the gradient $\nabla f(u) \in \mathbb{U}$ as the Riesz representation of the Fréchet derivative $\mathscr{D}f(u) \in \mathbb{U}^*$.

With Definition 9.3 at hand, we can identify the gradient alias of the Fréchet derivative and we will explore this fact in many results below.

Example 9.1. Consider $f : \mathbb{U} \to \mathbb{R}$ where $\mathbb{U} \equiv \mathbb{R}^n$ and $\mathbb{R}^n$ is endowed with a weighted inner product $(\boldsymbol{u}, \boldsymbol{v})_{\mathbb{R}^n, \boldsymbol{M}} = \boldsymbol{u}^T \boldsymbol{M} \boldsymbol{v}$ with $\boldsymbol{M}$ being a symmetric positive definite matrix. Suppose that the (classical) partial derivatives $\dfrac{\partial f}{\partial \boldsymbol{u}_i}$, $i = 1, \ldots, n$, of the f are continuous. From (9.2), it is easy to see that the Fréchet derivative can be written as

$$\mathscr{D}f(\boldsymbol{u}, \boldsymbol{h}) = \sum_{i=1}^{N} \frac{\partial f}{\partial \boldsymbol{u}_i} h_i = \left[\frac{\partial f}{\partial \boldsymbol{u}_1}, \ldots, \frac{\partial f}{\partial \boldsymbol{u}_n} \right]^T \boldsymbol{h},$$

which, together with Definition 9.3, gives

$$\nabla f(\boldsymbol{u}) = \boldsymbol{M}^{-1} \left[\frac{\partial f}{\partial \boldsymbol{u}_1}, \ldots, \frac{\partial f}{\partial \boldsymbol{u}_n} \right]^T.$$

We observe that the Fréchet derivative is a special case of the directional derivative, and when $\boldsymbol{M}$ is the identity matrix, the classical gradient vector $\left[\dfrac{\partial f}{\partial \boldsymbol{u}_1}, \ldots, \dfrac{\partial f}{\partial \boldsymbol{u}_n} \right]^T$ is in fact the Riesz representation of the Fréchet derivative in the standard Euclidean inner product.

Of course, the Fréchet derivative can be directly generalized to mappings between two different vector spaces. For example, if $c : \mathbb{U} \ni u \mapsto c(u) \in \mathbb{V}$, then the Fréchet derivative $\mathscr{D}c(u)$, when exists, can be computed as

$$\mathscr{D}c(u)\, v := \mathscr{D}c(u; v) := \lim_{t \to 0} \frac{c(u + tv) - c(u)}{t}.$$

The difference is now that $\mathscr{D}c(u)$ is a linear and bounded map from $\mathbb{U}$ to $\mathbb{V}$, that is, $\mathscr{D}c(u) \in \mathcal{B}(\mathbb{U}, \mathbb{V})$.

Example 9.2. Consider a vector-valued function $\mathbf{c}(\boldsymbol{u}) : \mathbb{R}^n \to \mathbb{R}^m$ where both $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the standard Euclidean inner products. Applying Example 9.1 for each component of $\mathbf{c}_i$, $i = 1, \ldots, m$ we have

$$\mathscr{D}\mathbf{c}(\boldsymbol{u})\, \boldsymbol{v} = \begin{bmatrix} \dfrac{\partial \mathbf{c}_1}{\partial \boldsymbol{u}_1} & \dfrac{\partial \mathbf{c}_1}{\partial \boldsymbol{u}_2} & \cdots & \dfrac{\partial \mathbf{c}_1}{\partial \boldsymbol{u}_n} \\ \dfrac{\partial \mathbf{c}_2}{\partial \boldsymbol{u}_1} & \dfrac{\partial \mathbf{c}_2}{\partial \boldsymbol{u}_2} & \cdots & \dfrac{\partial \mathbf{c}_2}{\partial \boldsymbol{u}_n} \\ \vdots & \vdots & \cdots & \vdots \\ \dfrac{\partial \mathbf{c}_m}{\partial \boldsymbol{u}_1} & \dfrac{\partial \mathbf{c}_m}{\partial \boldsymbol{u}_2} & \cdots & \dfrac{\partial \mathbf{c}_m}{\partial \boldsymbol{u}_n} \end{bmatrix} \boldsymbol{v},$$

which, together with Definition 9.3, we can define

$$\nabla \boldsymbol{c}(\boldsymbol{u}) := \begin{bmatrix} \frac{\partial \boldsymbol{c}_1}{\partial \boldsymbol{u}_1} & \frac{\partial \boldsymbol{c}_1}{\partial \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{c}_1}{\partial \boldsymbol{u}_n} \\ \frac{\partial \boldsymbol{c}_2}{\partial \boldsymbol{u}_1} & \frac{\partial \boldsymbol{c}_2}{\partial \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{c}_2}{\partial \boldsymbol{u}_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \boldsymbol{c}_m}{\partial \boldsymbol{u}_1} & \frac{\partial \boldsymbol{c}_m}{\partial \boldsymbol{u}_2} & \cdots & \frac{\partial \boldsymbol{c}_m}{\partial \boldsymbol{u}_n} \end{bmatrix},$$

which is the Riesz representation of $\mathscr{D}\boldsymbol{c}(\boldsymbol{u})$.

Lemma 9.2 (First order optimality condition for unconstrained optimization). *Suppose that $f : \mathbb{U} \to \mathbb{R}$ attains its extremum at u. Then it is necessary that*

$$\mathscr{D}f(u)\, v = 0, \quad \forall v \in \mathbb{U}, \tag{9.3}$$

that is, the (first) variation of f at u in any "direction" v vanishes. In other words, it is necessary that $\mathscr{D}f(u) = 0$ or equivalently

$$\nabla f(u) = 0,$$

by the Riesz representation Theorem 5.1.

Proof. It is sufficient to assume that f is minimized at u, i.e.,

$$f(v) \geq f(u), \quad \forall v \in \boldsymbol{B}_\delta(u) := \{w \in \mathbb{U} : \|w - u\|_{\mathbb{U}} < \delta\},$$

which implies that for any v such that $\varepsilon \|v\|_{\mathbb{U}} < \delta$, we have

$$f(u + \varepsilon v) \geq f(u), \quad \forall \varepsilon \in \boldsymbol{B}_{\delta/\|v\|_{\mathbb{U}}}(0). \tag{9.4}$$

If we define $F(\varepsilon) := f(u + \varepsilon v)$, then $F(\cdot)$ is a function in ε, namely, $F : \mathbb{R} \ni \varepsilon \mapsto F(\varepsilon) \in \mathbb{R}$. By Definition 9.1, inequality (9.4) is equivalent to saying that $F(\cdot)$ attains its minimum at $\varepsilon = 0$. Thus, from the first result of Lemma 9.1, we have

$$\left.\frac{dF}{d\varepsilon}\right|_{\varepsilon=0} = 0,$$

but this is equivalent to $\mathscr{D}f(u, v) = 0$ by Definition 9.2 of Fréchet derivative.

Example 9.3 (First order optimality condition for unconstrained optimization in $\mathbb{R}^n$). Back to Example 9.1. Suppose that f attains its minimum at $\boldsymbol{u}$. Combining (9.3), the gradient found in Example 9.1, and Definition 9.3 yields

$$\nabla f(\boldsymbol{u}) = \boldsymbol{0},$$

and thus the first-order necessary condition for optimality is given by

$$\left[\frac{\partial f}{\partial \boldsymbol{u}_1}, \ldots, \frac{\partial f}{\partial \boldsymbol{u}_n}\right]^T = \mathbf{0}.$$

Example 9.4. We now revisit the least squares problem in Corollary 11.1 in the equivalent form: $\inf_{u \in \mathbb{U}} \frac{1}{2} (\mathscr{A}u - v, \mathscr{A}u - v)_{\mathbb{V}}$. Using Definition 9.2, the first order optimality condition (9.3) reads

$$2 (\mathscr{A}v, \mathscr{A}u - v)_{\mathbb{V}} = 0, \quad \forall v \in \mathbb{U},$$

that is,

$$\mathscr{A}^* \mathscr{A} u = \mathscr{A}^* v,$$

which is consistent with the least squares solution in (11.1).

9.3 First-order optimality conditions for optimization problems with equality constraints in any dimensions

Up to this point, we have looked at unconstrained optimization problems and derived the (first order) necessary condition for optimality. We next discuss optimality conditions for constrained optimization. Let us consider the following constrained optimization problem

$$\min_{u \in \mathbb{U}} f(u), \qquad \text{subject to } c(u) = \theta, \text{ where } c(\cdot) : \mathbb{U} \to \mathbb{V}.$$

If there were no constraint $c(u) = \theta$, then from Lemma 9.2 the optimality condition would be

$$\mathscr{D}f(u)\, v = 0, \quad \forall v \in \mathbb{U},$$

That is, the variation of f at u in any "direction" v vanishes. However, v can be no longer arbitrary since the constraint must be satisfied at $u + tv$ for any small t. In other words, $u + tv$ needs to be *feasible*, i.e.,

$$c(u + tv) = 0, \qquad \text{for any feasible } u + tv.$$

Suppose c is Fréchet differentiable, it is therefore necessary that

$$\mathscr{D}c(u)\, v = 0, \text{ for any feasible } u + tv.$$

To rigorously establish this result, we need the inverse function theorem [97], which in turn is a direct consequence of the implicit function theorem [88, 50].

Theorem 9.1 (Implicit function theorem). *Let $c : \mathbb{U} \times \mathbb{Z} \to \mathbb{V}$ be continuously Fréchet differentiable and $\mathscr{D}_u c(u, z) : \mathbb{U} \to \mathbb{V}$, is invertible at a point*

$[u_0, z_0]^T$, *at which* $c(u_0, z_0) = 0$. *Then, there exist a neighborhood* $\boldsymbol{B}_\delta(z_0)$ *and a continuously Fréchet differential function* $g : \boldsymbol{B}_\delta(z_0) \to \mathbb{U}$ *such that* $c(g(z), z) = 0$ *for all* $z \in \boldsymbol{B}_\delta(z_0)$.

Theorem 9.2 (Inverse function theorem). *Let* $f : \mathbb{U} \to \mathbb{Z}$. *Assume that* $\mathscr{D}f(u_0)$ *is continuous and maps* $\mathbb{U}$ **onto** $\mathbb{Z}$. *Then, there is a neighborhood* $\boldsymbol{B}_\delta(f(u_0))$ *of* $f(u_0)$ *such that* $f(u) = z$ *has a unique continuously differentiable solution* $u(z)$ *for every* $z \in \boldsymbol{B}_\delta(f(u_0))$.

Proof. The result is clear if we define $c(u, z) = z - f(u)$ and set $z_0 = f(u_0)$. Then by the implicit function Theorem 9.1, there exists $g : \boldsymbol{B}_\delta(z_0) \to \mathbb{U}$ such that $0 = c(g(z), z) = z - f(g(z)) = 0$ for all $z \in \boldsymbol{B}_\delta(z_0)$. Setting $u = g(z)$ concludes the proof.

Lemma 9.3 (First order optimality condition for equality constraints). *Suppose* $f : \mathbb{U} \to \mathbb{R}$ *attains its extremum at* u_0 *subject to the constraint* $c(u) = 0$, *where* $c : \mathbb{U} \to \mathbb{V}$. *Assume that both* f *and* c *are continuously Fréchet differentiable in an open set containing* u_0, *and* $\mathscr{D}c(u_o)$ *maps* $\mathbb{U}$ **onto** $\mathbb{V}$. *Then, it is necessary that*

$$\mathscr{D}f(u_0)\, v = 0, \quad \forall v \in \mathbb{U} \text{ such that } \mathscr{D}c(u_0)\, v = \theta,$$

or equivalently

$$(\nabla f(u_0), v)_{\mathbb{U}} = 0, \quad \forall v \in \mathbb{U} \text{ such that } \mathscr{D}c(u_0)\, v = \theta.$$

Proof. We follow closely the proof by contradiction in [97, Lemma 1 of Chapter 9]. Without lost of generality, asssume u_0 is a minimizer. Let us consider the transformation $g(u) = (f(u), c(u)) : \mathbb{U} \to \mathbb{R} \times \mathbb{V}$. Assume that there exists h such that $\mathscr{D}c(u_0; h) = 0$ but $\mathscr{D}f(u_0; h) \neq 0$. Then the function $(\mathscr{D}f(u_0), \mathscr{D}c(u_0))$ maps $\mathbb{U}$ onto $\mathbb{R} \times \mathbb{V}$ since $\mathscr{D}c(u_0)$ maps $\mathbb{U}$ onto $\mathbb{V}$. By the inverse function Theorem 9.2 there exists ε and u with $\|u - u_0\|_{\mathbb{U}} < \varepsilon$ such that $g(u) = (f(u_0) - \delta, 0)$ for some small $\delta > 0$. Thus, $f(u) = f(u_0) - \delta < f(u_0)$: a contradiction.

Example 9.5. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $\boldsymbol{A} \in \mathbb{R}^{m \times n}$, where $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the standard Euclidean inner products. We consider the following problem

$$\min_{\boldsymbol{u} \in \mathbb{R}^n} f(\boldsymbol{u}), \qquad \text{subject to } \boldsymbol{A}\boldsymbol{u} = \boldsymbol{b}.$$

From Lemma 9.3, Example 9.1, and Example 9.2 we can write the the first order optimality condition as

$$\nabla f^T(\boldsymbol{u})\, \boldsymbol{v} = 0, \quad \forall \boldsymbol{v} \in \mathbb{R}^n \text{ such that } \boldsymbol{A}\boldsymbol{v} = \boldsymbol{0},$$

or equivalently

$$\nabla f^T(\boldsymbol{u})\,\boldsymbol{v} = 0, \quad \forall \boldsymbol{v} \in \mathsf{N}(\boldsymbol{A}),$$

i.e., *due to the constraint, the gradient of f at an optimum $\boldsymbol{u}$ does not vanish but is orthogonal to the nullspace of the gradient $\boldsymbol{A}$ of the constraint.* In other words, for constrained optimization problems, at an optimum the projection of the gradient of the objective function in the nullspace of the gradient of the constraints vanishes. If we define $\boldsymbol{Z}$ with columns comprising a basis of the nullspace of $\boldsymbol{A}$, then $\boldsymbol{v} = \boldsymbol{Z}\mathbf{r}$ for some vector $\mathbf{r}$ whose dimension is the dimension of the nullspace. As a result, the constraint is completely eliminated and the first order optimality condition now reads

$$\nabla f^T(\boldsymbol{u})\,\boldsymbol{Z} = 0.$$

Note that $\mathbf{g}_r(\boldsymbol{u}) := \nabla f^T(\boldsymbol{u})\,\boldsymbol{Z}$—the coordinates of the gradient in the nullspace of the constraint gradient—is known as the reduced gradient [57, 2]. The reduced gradient is nothing more than the total gradient of the objective function with respect to the reduced variable $\mathbf{r}$ as we show below. The proof for an important class of constrained optimization is presented in Remark 9.3. One of the reasons for its name is that its dimension is smaller than the dimension n of the original gradient vector $\nabla f(\boldsymbol{u})$. *Another important point one can draw from the optimality condition for the reduced gradient is that in the reduced optimization variables $\mathbf{r}$, the optimization problem becomes implicitly unconstrained* (see the explicit transformation to the reduced space at the end of the example). Now from Corollary 9.1 we know that the derivative of the reduced gradient $\mathbf{g}_r$, namely the reduced Hessian $\boldsymbol{H}_r$, is necessary to be semi-positive definite in any direction in the reduced space. By the chain rule we have

$$\mathbf{r}^T \boldsymbol{H}_r(\boldsymbol{u})\,\mathbf{r} = \left.\frac{d\mathbf{g}_r(\boldsymbol{u} + tZ\mathbf{r})}{dt}\right|_{t=0} = \mathbf{r}^T \boldsymbol{Z}^T \nabla f^2(\boldsymbol{u})\,\boldsymbol{Z}\mathbf{r}, \quad \forall \mathbf{r},$$

from which it follows that

$$\boldsymbol{H}_r(\boldsymbol{u}) = \boldsymbol{Z}^T \nabla f^2(\boldsymbol{u})\,\boldsymbol{Z}.$$

Note that if QR factorization of $\boldsymbol{A}$ is feasible, then $\boldsymbol{Z}$ can be easily found. $\boldsymbol{Z}$ can also be explicitly identified for the case when $m \leq n$ and $\boldsymbol{A}$ has linearly independent rows. Indeed, up to a permutation of columns, we can rewrite $\boldsymbol{A}$ as

$$\boldsymbol{A} = [\boldsymbol{U}, \boldsymbol{V}],$$

where $\boldsymbol{U} \in \mathbb{R}^{m \times m}$ is a nonsingular matrix. Then $\boldsymbol{Z}$ can be written as

$$\boldsymbol{Z} = \begin{bmatrix} -\boldsymbol{U}^{-1}\boldsymbol{V} \\ I \end{bmatrix},$$

where I is the $(n-m)\times(n-m)$ identity matrix. In this case, we can write $\boldsymbol{u}=[\boldsymbol{u_U},\boldsymbol{u_V}]^T$ and we can eliminate $\boldsymbol{u_U}$ from the constraint as $\boldsymbol{u_U}=\boldsymbol{U}^{-1}\boldsymbol{b}-\boldsymbol{U}^{-1}\boldsymbol{V}\boldsymbol{u_V}$. The reduced optimization variable is thus $\boldsymbol{u_V}$ and the original constraint optimization problem is now unconstrained with respect to $\boldsymbol{u_V}$.

In order to provide further insights and make the optimality condition practical for large-scale computation we need a Lagrangian formalism, and *this is where the adjoint plays the key role*. Thanks to the closed range Theorem 5.2, the Lagrangian multiplier theorem [97, 19, 111, 18] is a straightforward equivalence to Lemma 9.3.

Theorem 9.3 (Lagrangian multiplier theorem). *Assume that $f:\mathbb{U}\to\mathbb{R}$ is continuously Fréchet differentiable and it attains the extremum at u_0 subject to the constraint $c(u)=0$, where $c:\mathbb{U}\to\mathbb{V}$ is continuously Fréchet differentiable. Suppose that $\mathscr{D}c(u_0)$ maps $\mathbb{U}$* **onto** *$\mathbb{V}$. Then, there exists an element $v\in\mathbb{V}$ such that the following Lagrangian functional*

$$L(u):=f(u)+(v,c(u))_{\mathbb{V}}$$

is stationary at u_0, i.e.,

$$\mathscr{D}L(u_0)\,h=\mathscr{D}f(u_0)\,h+(v,\mathscr{D}c(u_0)\,h)_{\mathbb{V}}=0,\quad \forall h\in\mathbb{U}, \tag{9.5}$$

or equivalently

$$\nabla L(u_0)=\nabla f(u_0)+[\mathscr{D}c(u_0)]^*\,v=0. \tag{9.6}$$

Proof. From Lemma 9.3 we see that $\nabla f(u_0)$ is orthogonal to the nullspace of $\mathscr{D}c(u_0)$. Since $\mathscr{D}c(u_0)$ maps $\mathbb{U}$ onto $\mathbb{V}$, the range of $\mathscr{D}c(u_0)$ closed and the closed range Theorem 5.2 gives

$$\nabla f(u_0)\in\mathsf{R}\left([\mathscr{D}c(u_0)]^*\right),$$

which implies that there exists $y\in\mathbb{V}$ such that

$$\nabla f(u_0)=-[\mathscr{D}c(u_0)]^*\,v,$$

and this ends the proof.

Remark 9.1. The appealing feature of the Lagrangian approach in Theorem 9.3 is that the first order optimality condition (9.6) is the standard optimality condition for unconstrained problem in Lemma 9.2, but for the Lagrangian instead of the original objective function f. The key implication in the Langrangian formalism is thus the optimization problem is unconstrained in the original optimization variable u plus the Lagrange multiplier v, as far as the first order optimality condition is concerned. *The Lagrangian approach is also known as the adjoint approach as it involves the adjoint of*

the gradient of the constraint in the first order optimality condition (9.6). If further structures of the constraints and/or optimization variables are given, the Lagrangian approach can lead to an efficient reduced space approach as we will discuss in the below examples, including the derivation and insights into backpropagation of neural network in Chapter 10.

Remark 9.2. Either (9.5) or (9.6) is know as the adjoint equation.

Example 9.6. Back to Example 9.5. Applying the Lagrangian multiplier Theorem 9.3 together with the Riesz representation Theorem 5.1, the equivalent first order optimality condition (9.6) reduces to

$$\nabla f(\boldsymbol{u}) + \boldsymbol{A}^T \boldsymbol{v} = 0,$$

which says that *at an optimum of a constrained optimization problem, the gradient of the objective function $f(\boldsymbol{u})$ does not vanish but is a linear combination of the gradient of the constraints.* Again, this is the same as saying that the gradient of the objective function $f(\boldsymbol{u})$ at an optimum is orthogonal to the nullspace of the gradient of the constraints.

Example 9.7. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{c}(\boldsymbol{u}) : \mathbb{R}^n \to \mathbb{R}^m$, where $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the standard Euclidean inner products. We consider the following optimization problem

$$\min_{\boldsymbol{u}\in\mathbb{R}^n} f(\boldsymbol{u}), \qquad \text{subject to } \mathbf{c}(\boldsymbol{u}) = \boldsymbol{b}.$$

From Lemma 9.3, Example 9.1, Example 9.2, and the Lagrangian multiplier Theorem 9.3 we can write the first order optimality condition either as

$$\nabla f(\boldsymbol{u}) + \nabla \mathbf{c}^T(\boldsymbol{u})\, \boldsymbol{v} = \mathbf{0}, \tag{9.7}$$

or as

$$\mathbf{g}(\boldsymbol{u}) := \nabla f^T(\boldsymbol{u})\, \boldsymbol{v}(\boldsymbol{u}) = 0, \quad \forall \boldsymbol{v}(\boldsymbol{u}) \in \mathbb{R}^n \text{ such that } \nabla \mathbf{c}(\boldsymbol{u})\, \boldsymbol{v}(\boldsymbol{u}) = \mathbf{0}.$$

Note that unlike the linear problem, $\boldsymbol{v}$ is a function of $\boldsymbol{u}$ since the nullspace of $\nabla \mathbf{c}(\boldsymbol{u})$ depends on $\boldsymbol{u}$. As a result, *the reduced Hessian is different from that of the linear constraint counterpart in Example 9.5 as we now show.* To that end, we compute the Fréchet derivative of $\mathbf{g}$ in any direction $\mathbf{p}$ in the reduced space $\mathsf{N}(\nabla \mathbf{c}(\boldsymbol{u}))$:

$$\mathbf{p}^T \boldsymbol{H} \mathbf{h} = \mathbf{p}^T \nabla f^2 \boldsymbol{v} + \nabla f^T \mathscr{D}\boldsymbol{v}(\boldsymbol{u}, \mathbf{p}). \tag{9.8}$$

To compute $\mathscr{D}\boldsymbol{v}(\boldsymbol{u}, \mathbf{p})$ we take the Fréchet derivative both sides of $\nabla \mathbf{c}(\boldsymbol{u})\, \boldsymbol{v}(\boldsymbol{u}) = \mathbf{0}$, row by row, in direction $\mathbf{p}$ to arrive at

$$\mathbf{p}^T \nabla^2 c_i \boldsymbol{v} + \nabla c_i^T \mathscr{D}\boldsymbol{v}(\boldsymbol{u}, \mathbf{p}) = 0. \tag{9.9}$$

Combining (9.7)–(9.9) gives

$$\mathbf{p}^T \boldsymbol{H}\mathbf{h} = \mathbf{p}^T \left(\nabla f^2 - \sum_{i=1}^{m} \nabla^2 c_i y_i \right) \boldsymbol{v}.$$

Since both $\mathbf{p}$ and $\boldsymbol{v}$ belong to the nullspace $\boldsymbol{Z}(\boldsymbol{u})$ of $\nabla \mathbf{c}(\boldsymbol{u})$, the reduced Hessian $\boldsymbol{H}_r$ is then given by

$$\boldsymbol{H}_r = \boldsymbol{Z}^T \left(\nabla f^2 - \sum_{i=1}^{m} \nabla^2 c_i y_i \right) \boldsymbol{Z}.$$

As can be seen, the Hessians of the constraints (which is zero for linear constraint case in Example 9.5) contribute to the reduced Hessian. Unlike Example 9.5 in which the reduced space, and hence the reduced optimization variables, is fixed if an iterative gradient-based algorithm is employed, the reduced space for this example changes at each optimization step due the nonlinear nature of the constraint $\mathbf{c}(\boldsymbol{u}) = \boldsymbol{b}$.

9.4 The reduced-space approach for separable optimization problems with equality constraints

In this section, we consider an important class of constrained optimization problems in which the constraints are equalities and the optimization variables are separable in the sense that from the constraint one can solve for one sub-variable as a function of the other. As we shall see in Chapter 10, training deep neural network with backpropagation is a special case of this class. PDE-constrained optimization problem is another special case as shown in Chapter 17.

Corollary 9.2 (Optimization with special equality constraint). *Consider optimization problems that can be expressed in the following form*

$$\min_{u \in \mathbb{U}, z \in \mathbb{Z}} f(u, z), \qquad \textit{subject to } \; c(u, z) = 0, \; \textit{where } c(\cdot, \cdot) : \mathbb{U} \times \mathbb{Z} \to \mathbb{V},$$

where the Fréchet derivative of the constraint with respect to u, i.e. $\mathscr{D}_u c(u, z) : \mathbb{U} \to \mathbb{V}$, is invertible *at an optimum* $[u_0, z_0]$. *The first order optimality condition (9.6), together with the constraint, can be written as*

$$c(u_0, z_0) = 0, \qquad \textit{Forward equation}, \tag{9.10a}$$

$$\nabla_u f(u_0, z_0) + [\mathscr{D}_u c(u_0, z_0)]^* y = 0, \qquad \textit{Adjoint equation}, \tag{9.10b}$$

$$\nabla_z f(u_0, z_0) + [\mathscr{D}_z c(u_0, z_0)]^* y = 0, \qquad \textit{Control equation}, \tag{9.10c}$$

where ∇_u and ∇_z denote the Fréchet derivative with respect to u and z, respectively.

Proof. The proof is a straightforward application of (9.6) to the group optimization variable $[u, z]^T$, and thus $\nabla f = [\nabla_u f, \nabla_z f]^T$ and $\mathscr{D}c = [\mathscr{D}_u c, \mathscr{D}_z c]^T$.

Note that the left-hand side (LHS) of the forward problem (9.10a) is simply the derivative of the Lagrangian with respect to the adjoint variable v. The LHS of the adjoint equation (9.10b) is nothing more than the derivative of the Lagrangian with respect to u, and the LHS of the control equation (9.10c) is the derivative of the Lagrangian with respect to z. For this class of optimization problems, we can eliminate both the "state" variable u and adjoint variable v so that the optimization problem is genuinely unconstrained in only the control variable z around a neighborhood of the optimum $[u_0, z_0]^T$. Indeed, from the implicit function Theorem 9.1 there exists $\boldsymbol{B}_\delta(z_0)$ and a continuously differentiable function $g : \boldsymbol{B}_\delta(z_0) \to \mathbb{U}$ such that $u = g(z)$ solve the constraint $c(u, z) = 0$ for any $z \in \boldsymbol{B}_\delta(z_0)$. The objective function becomes $f(g(z), z)$, and thus a function of only z, for all $z \in \boldsymbol{B}_\delta(z_0)$. The optimization variable is reduced to only z and this is known as the *reduced space approach* [57, 2]. Clearly, we do not know $u = g(z)$ explicitly, and the question is how to compute the reduced gradient $\nabla f(g(z), z)$, which is needed for any gradient-based approach? Note that by $\nabla f(g(z), z)$ we mean the total derivative with respect to z.

Lemma 9.4 (Reduced gradient in constrained optimization with equality constraints). *With the same setting as in Corollary 9.2, there exists a neighborhood $\boldsymbol{B}_\delta(z_0)$ such that the reduced gradient at any $z \in \boldsymbol{B}_\delta(z_0)$ is given by*

$$\nabla f(g(z), z) = \nabla_z f(u, z) + [\mathscr{D}_z c(u, z)]^* y, \tag{9.11}$$

where $u = g(z)$ and v satisfy the following forward and adjoint equations:

$$c(u, z) = 0, \qquad \textit{Forward equation}, \tag{9.12a}$$

$$\nabla_u f(u, z) + [\mathscr{D}_u c(u, z)]^* y = 0, \qquad \textit{Adjoint equation}. \tag{9.12b}$$

Proof. Note that the invertibility of a linear operator $\mathscr{A}$ implies the invertibility[2] of its adjoint $\mathscr{A}^*$ with $(\mathscr{A}^*)^{-1} = (\mathscr{A}^{-1})^*$. For simplicity, we use $(\mathscr{A})^{-*}$ to denote $(\mathscr{A}^{-1})^*$. We have

[2] Indeed, suppose $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ is invertible, then $\langle u, \mathscr{A}^* (\mathscr{A}^{-1})^* w \rangle_{\mathbb{U}} = \langle \mathscr{A}u, (\mathscr{A}^{-1})^* w \rangle_{\mathbb{V}} = \langle \mathscr{A}^{-1}\mathscr{A}u, w \rangle_{\mathbb{U}} = \langle u, w \rangle_{\mathbb{U}}$ for any $u, w \in \mathbb{U}$. Thus, $(\mathscr{A}^*)^{-1} = (\mathscr{A}^{-1})^*$.

$$
\begin{aligned}
&\left(\nabla f\left(g\left(z\right),z\right),h\right)_{\mathbb{Z}} \\
&\quad\| && \text{Gradient Definition 9.3} \\
&\left\langle \mathscr{D}f\left(g\left(z\right),z\right),h\right\rangle_{\mathbb{Z}} \\
&\quad\| && \text{Chain rule} \\
&\left\langle \mathscr{D}_z f\left(u,z\right),h\right\rangle_{\mathbb{Z}}+\left\langle \mathscr{D}_u f\left(u,z\right),\mathscr{D}_z g\left(z\right)h\right\rangle_{\mathbb{U}} \\
&\quad\| && \text{Gradient Definition 9.3} \\
&\left(\nabla_z f\left(u,z\right),h\right)_{\mathbb{Z}}+\left(\nabla_u f\left(u,z\right),\mathscr{D}_z g\left(z\right)h\right)_{\mathbb{U}} \\
&\quad\| && \text{Derivative of the constraint} \\
&\left(\nabla_z f\left(u,z\right),h\right)_{\mathbb{Z}}-\left(\nabla_u f\left(u,z\right),\left[\mathscr{D}_u c\left(u,z\right)\right]^{-1}\mathscr{D}_z c\left(u,z\right)h\right)_{\mathbb{U}} \\
&\quad\| && \text{Adjoint Definition 5.6} \\
&\left(\nabla_z f\left(u,z\right),h\right)_{\mathbb{Z}}-\left(\left[\mathscr{D}_z c\left(u,z\right)\right]^{*}\left[\mathscr{D}_u c\left(u,z\right)\right]^{-*}\nabla_u f\left(u,z\right),h\right)_{\mathbb{Z}} \\
&\quad\| && \left[\mathscr{D}_u c\left(u,z\right)\right]^{*}v:=-\nabla_u f\left(u,z\right) \\
&\left(\nabla_z f\left(u,z\right),h\right)_{\mathbb{Z}}+\left(\left[\mathscr{D}_z c\left(u,z\right)\right]^{*}y,h\right)_{\mathbb{Z}}
\end{aligned}
$$

where, as in the first equality, the derivative of the constraint in the third equality is given by the chain rule[3]:

$$\mathscr{D}_z c\left(u,z\right)+\mathscr{D}_u c\left(u,z\right)\mathscr{D}_z g\left(z\right)=\theta,$$

which, due to the invertibility of $\mathscr{D}_u c\left(u,z\right)$, allows us to solve for $\mathscr{D}_z g\left(z\right)$.

Remark 9.3. Note that in practice, approximating the minimum u_0 is typically done using a gradient descent algorithm and Lemma 9.4 shows that the reduced gradient can be computed in each iteration for a given z via three steps: 1) solve the *forward equation* (9.12a) for $u\left(z\right)$, 2) solve the *adjoint equation* (9.12b) for $y\left(u(z),z\right)$, and 3) substitute $u(z)$ and $v(u\left(z\right),z)$ into (9.11) to obtain the reduced gradient. Moreover, the adjoint equation is always linear in the adjoint variable v regardless the linear or nonlinear nature of the forward equation. Note that full space iteration based on the first order optimality condition (9.10) is also possible and can be consulted from [111] and the references therein.

Example 9.8. To appreciate the adjoint approach, let us apply Lemma 9.4 to identify the forward equation, the adjoint equation, and the reduced gradient of the following finite dimensional optimization problem

$$\min_{\boldsymbol{u}\in\mathbb{R}^n,\mathbf{z}\in\mathbb{R}^p} f\left(\boldsymbol{u},\mathbf{z}\right),\qquad \text{subject to }\ \mathbf{c}\left(\boldsymbol{u},\mathbf{z}\right)=\mathbf{0},$$

where $f:\mathbb{R}^n\times\mathbb{R}^p\to\mathbb{R}$ and $\mathbf{c}\left(\boldsymbol{u},\mathbf{z}\right):\mathbb{R}^n\times\mathbb{R}^p\to\mathbb{R}^n$, and all spaces are equipped with the standard Euclidean inner products. We assume that

[3] The chain rule for Fréchet derivation can be derived from (9.2). Let $f:\mathbb{U}\ni u\mapsto f\left(u\right)\in\mathbb{V}$ and $g:\mathbb{Z}\ni z\mapsto g\left(z\right)\in\mathbb{U}$. We have $f\left(g\left(z+\varepsilon h\right)\right)=f\left(g\left(z\right)+\varepsilon\mathscr{D}g\left(z\right)h+o\left(\varepsilon\right)\right)=f\left(g\left(z\right)\right)+\varepsilon\mathscr{D}_u f\left(g\left(z\right)\right)\left(\mathscr{D}g\left(z\right)h+\varepsilon^{-1}o\left(\varepsilon\right)\right)+o\left(\varepsilon\right)$. Thus, $\mathscr{D}_z f\left(g\left(z\right)\right)h=\lim_{\varepsilon\to 0}\frac{f\left(g\left(z+\varepsilon h\right)\right)-f\left(g\left(z\right)\right)}{\varepsilon}=\mathscr{D}_u f\left(g\left(z\right)\right)\mathscr{D}g\left(z\right)h=\left\langle \mathscr{D}_u f\left(g\left(z\right)\right),\mathscr{D}g\left(z\right)h\right\rangle_{\mathbb{U}}$.

$\det(\nabla_{\boldsymbol{u}}\mathbf{c}) \neq 0, \forall \boldsymbol{u}, \mathbf{z}$ so that the implicit function theorem allows us to compute $\boldsymbol{u}$ as a function of $\mathbf{z}$ from the constraint. Applying Lemma 9.4 the reduced gradient reads

$$\nabla f = \nabla_{\mathbf{z}} f(\boldsymbol{u}_0, \mathbf{z}_0) + [\nabla_{\mathbf{z}}\mathbf{c}(\boldsymbol{u}_0, \mathbf{z}_0)]^T \boldsymbol{v}, \tag{9.13}$$

where $\boldsymbol{u}$ and $\boldsymbol{v}$ are computed from

$$\mathbf{c}(\boldsymbol{u}_0, \mathbf{z}_0) = \mathbf{0}, \qquad \text{Forward equation,} \tag{9.14a}$$

$$\nabla_{\boldsymbol{u}} f(\boldsymbol{u}_0, \mathbf{z}_0) + [\nabla_{\boldsymbol{u}}\mathbf{c}(\boldsymbol{u}_0, \mathbf{z}_0)]^T \boldsymbol{v} = \mathbf{0}, \qquad \text{Adjoint equation.} \tag{9.14b}$$

Example 9.9. For a more concrete example, let us apply Lemma 9.4 to identify the forward equation, the adjoint equation, and the reduced gradient of the following simplified version of Example 9.8

$$\min_{\boldsymbol{u}\in\mathbb{R}^n, \mathbf{z}\in\mathbb{R}^p} J = \frac{1}{2}\left\|\boldsymbol{f}^{obs} - \boldsymbol{C}\boldsymbol{u}\right\|^2_{\mathbb{R}^m}, \quad \text{subject to } \boldsymbol{A}\boldsymbol{u} + \boldsymbol{B}\mathbf{z} = \mathbf{0},$$

where $\boldsymbol{C} \in \mathbb{R}^{m\times n}$, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is invertible, and $\boldsymbol{B} \in \mathbb{R}^{n\times p}$. The reduced gradient in (9.13) simplifies to

$$\nabla J = \boldsymbol{B}^T \boldsymbol{v},$$

and the system (9.14) reduces to

$$\begin{aligned} \boldsymbol{A}\boldsymbol{u} + \boldsymbol{B}\mathbf{z} &= \mathbf{0}, , \\ -\boldsymbol{C}^T\left(\boldsymbol{f}^{obs} - \boldsymbol{C}\boldsymbol{u}\right) + \boldsymbol{A}^T\boldsymbol{v} &= \mathbf{0}. \end{aligned}$$

Example 9.10. We consider an open and bounded domain $\Omega \subset \mathbb{R}^d$ and recall from Example 12.7 the $\mathbb{L}^2-$inner product of two functions $u(\boldsymbol{u}), v(\boldsymbol{u})$ in $\mathbb{L}^2(\Omega)$ as

$$(u, v)_{L^2(\Omega)} = \int_\Omega u(\boldsymbol{u}) v(\boldsymbol{u}) d\Omega, \tag{9.15}$$

with the induced norm $\|u\|_{L^2(\Omega)} = \sqrt{(u,u)_{L^2(\Omega)}}$ and

$$\mathbb{L}^2(\Omega) = \left\{ f : \Omega \to \mathbb{R} : \|f\|_{\mathbb{L}^2(\Omega)} < \infty \right\}.$$

Let $\varphi \in L^2(\Omega)$ and $k(\cdot,\cdot;\boldsymbol{\alpha}) : \Omega\times\Omega \to \mathbb{R}$ such that $\int_\Omega \int_\Omega |k(\boldsymbol{u},\boldsymbol{w};\boldsymbol{\alpha})|^2 \, d\boldsymbol{u}\, d\boldsymbol{w} < \infty$ for any $\boldsymbol{\alpha} \in \mathbb{R}^p$. Consider the following optimization problem[4]

$$\min_{\boldsymbol{\alpha}\in\mathbb{R}^p} J = \frac{1}{2}\left(f^{obs} - (\varphi, u)_{\mathbb{L}^2(\Omega)}\right)^2$$

$$\text{subject to } \; u(\boldsymbol{u}) + \int_\Omega k(\boldsymbol{u},\boldsymbol{w};\boldsymbol{\alpha})\, u(\boldsymbol{w}) \, d\boldsymbol{w} = g(\boldsymbol{u}),$$

where $g \in L^2(\Omega)$ is a given function, and f^{obs} is a given number.

Note that the constraint can be written as

$$(\mathscr{I} + \mathsf{K})\, u = g, \tag{9.16}$$

where $\mathscr{I} : L^2(\Omega) \to L^2(\Omega)$ is the identity operator and the operator $\mathsf{K} : L^2(\Omega) \to L^2(\Omega)$ is defined via

$$(\mathsf{K}u)(\boldsymbol{u}) = \int_\Omega k(\boldsymbol{u},\boldsymbol{w};\boldsymbol{\alpha})\, u(\boldsymbol{w}) \, d\boldsymbol{w},$$

which is a compact operator for each $\boldsymbol{\alpha}$ (see Chapter 14), and thus by the Riesz-Fredholm theory [36], (9.16) has a unique solution u for each g and each $\boldsymbol{\alpha}$. Thus, $\boldsymbol{\alpha}$ is the reduced optimization variable and we now apply Lemma 9.4 to obtain the reduced gradient

$$\nabla J = \int_\Omega v(\boldsymbol{u}) \int_\Omega \nabla_{\boldsymbol{\alpha}} k(\boldsymbol{u},\boldsymbol{w};\boldsymbol{\alpha})\, u(\boldsymbol{w}) \, d\boldsymbol{w}\, d\boldsymbol{u},$$

where u and v are governed by the forward and adjoint system (9.12), which now becomes

$$\begin{aligned} (\mathscr{I} + \mathsf{K})\, u &= g \\ -\left(f^{obs} - (\varphi, u)_{\mathbb{L}^2(\Omega)}\right)\varphi + (\mathscr{I} + \mathsf{K}^*)\, v &= 0, \end{aligned}$$

where

$$(\mathsf{K}^* v)(\boldsymbol{w}) := \int_\Omega k(\boldsymbol{u},\boldsymbol{w};\alpha)\, v(\boldsymbol{u}) \, d\boldsymbol{u}.$$

Note that at the current $\boldsymbol{\alpha}$ where we would like to compute the reduced gradient, both forward state u and adjoint state v can be computed uniquely, thanks to the compactness of K (and hence K^*) and the Riesz-Fredholm theory [36]. Some variants of this example can be found Problem 9.4.

[4] This is a formulation for inverse scattering problem via the integral equation constraint of Lippmann-Schwinger type [34].

Problems

Problem 9.1. Show that Fréchet derivative defined in (9.2) is unique.

Hint: Assume there is another Fréchet derivative $\hat{\mathscr{D}} f(u,\cdot)$, then

$$\left|\hat{\mathscr{D}} f(u,h) - \mathscr{D} f(u,h)\right| \leq |f(u+h) - f(u) - \mathscr{D} f(u,h)| + \left|f(u+h) - f(u) - \hat{\mathscr{D}} f(u,h)\right| = o\left(\|h\|_{\mathbb{U}}\right).$$

but both $\mathscr{D} f(u,\cdot)$ and $\hat{\mathscr{D}} f(u,\cdot)$ are linear and bounded functional, the LHS $= \mathcal{O}(h)$, and this can only be true if the LHS is identically zero for all h and this implies $\mathscr{D} f(u,\cdot) = \hat{\mathscr{D}} f(u,\cdot)$.

Problem 9.2. Recall that $f : \mathbb{U} \to \mathbb{F}$ is continuous at u if, $\forall \varepsilon > 0$, there exist $\delta > 0$ such that

$$\|v-u\|_{\mathbb{U}} < \delta \implies |f(v) - f(u)| < \varepsilon.$$

Show that if f is Fréchet differentiable at u, then it is continuous at u.

Problem 9.3. Let $\mathbb{U} = \mathbb{L}^2(0,1)$ with a weighted inner product $(u,v)_{\mathbb{U}} := \int_0^1 u(t)\, v(t)\, w(t)\, dt$, where $w(t)$ is a positive function, and $f : \mathbb{U} \to \mathbb{R}$ defined as $f(u) := \int_0^1 g(u(t),t)\, dt$. Assume that $\dfrac{\partial g}{\partial u}$ exists and is continuous with respect to u and t. Derive the expression for the Fréchet derivative at u acting on h and then obtain the corresponding Frćhet gradient.

Hint:

$$\mathscr{D} f(u,h) = \int_0^1 \frac{\partial g}{\partial u}(u,t)\, h(t)\, dt = \left(\frac{1}{w}\frac{\partial g}{\partial u}, h\right)_{\mathbb{U}}.$$

the Fréchet gradient is thus $\nabla f = \dfrac{1}{w}\dfrac{\partial g}{\partial u}$.

Problem 9.4. Derive the reduced gradient for the following modification of Example 9.10.

1. Consider the following objective function
$$J := (\varphi, u)_{\mathbb{L}^2(\Omega)}$$
2. Consider now $\alpha \in \mathbb{L}^2(\Omega)$ instead of $\boldsymbol{\alpha} \in \mathbb{R}^n$.
3. Combine the above two modifications.

Chapter 10
Adjoint approach as backpropagation for deep learning

Abstract In this chapter, we consider standard fully connected deep neural network and use the adjoint method in Example 9.8 to derive the backpropagation method for computing the gradient of the loss function with respect to the weights and biases of a general fully-connected deep neural work (DNN). Excellent review papers on deep learning can be found in [90, 128], and the history of back-propagation can be traced back to [93, 94]. The gradient is needed for gradient-based methods (see, e.g., [111] and the references therein) such as stochastic gradient descent [83, 121, 130]. The extension of the adjoint method for other type of neural networks such as ResNet [70] and CNN [60, 75] are straightforward. We are going to show that *the backpropagation is nothing more than a reduced space approach to compute the gradient using adjoint method.* This finding also facilitates efficient computation of higher derivatives for DNN and this is the focus of the final part of the chapter. It is useful for higher-order optimization approaches, such as Newton-type methods. We shall show how to compute the product of the Hessian with an arbitrary vector exactly without forming the Hessian. This is extremely useful for Krylov subspace approaches, such as conjugate gradient, in training large-scale DNNs. Clearly, one could invoke the automatic differentiability of `Jax` [24] or `TensorFlow` [1] or `PyTorch` [113] to compute the gradient and possibly the Hessian-vector products. However, our exposition not only provides insights into the backpropagation via adjoint, but also facilitates the readers to write (and thus save memory as computational graphs are never needed in this case) their own optimized gradient and Hessian bypassing the automatic differentiation entirely.

10.1 Backpropagation under adjoint lense

This section starts with a definition of an ℓ-layer deep neural network (DNN). The DNN training problem is posed as an optimization problem and the

gradient computation is then derived using the adjoint approach. The key finding is that *the backpropagation is thus nothing more than a reduced space approach to compute the gradient using adjoint method.* Another important consequence of this finding is that the Hessian-vector products (that would be needed for Newton-type methods with Krylov subspace methods such as conjugate gradient method) can be computed exactly using the same adjoint approach, and this will be the focus of the next section.

Definition 10.1 (ℓ-layer Neural network). Given $\ell, s_0, s_1, \ldots, s_\ell \in \mathbb{N}$, an ℓ-layer neural network is defined as the following series of composition

$$\begin{aligned} &\text{Input layer} : \boldsymbol{a}^0 - \boldsymbol{x} = \mathbf{0}, \\ &\text{The } i\text{th layer} : \boldsymbol{a}^i - \sigma\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right) = \mathbf{0}, \quad i = 1, \ldots, \ell, \end{aligned} \tag{10.1}$$

where $\boldsymbol{x} \in \mathbb{R}^{s_0}$; $\boldsymbol{W}^i \in \mathbb{R}^{s_i} \times \mathbb{R}^{s_{i-1}}$ and $\boldsymbol{b}^i \in \mathbb{R}^{s_i}$, $i = 1, \ldots, \ell$, are weight matrix and bias vector of the ith layer; $\boldsymbol{a}^i \in \mathbb{R}^{s_i}$ is the output of the ith layer; and the activation function, σ, **acts component-wise** when its argument is a vector.

Let us define $\boldsymbol{u} := \left[\boldsymbol{a}^0, \ldots, \boldsymbol{a}^\ell\right]^T$, $\mathbf{z} := \left[\boldsymbol{W}^1, \boldsymbol{b}^1, \ldots, \boldsymbol{W}^\ell, \boldsymbol{b}^\ell\right]^T$, and $\boldsymbol{c}(\boldsymbol{u}, \mathbf{z}) = \mathbf{0}$ as the concatenation of all the sub-equations in (10.1). For concreteness, let us consider the loss (objective) function to be:

$$J(\boldsymbol{u}, \mathbf{z}) = \frac{1}{2} \left\| \boldsymbol{a}^{obs} - \boldsymbol{a}^\ell \right\|^2, \tag{10.2}$$

where $\boldsymbol{a}^{obs}$ is a given data (label). The neural network training problem is exactly the constrained optimization problem in Example 9.8. Thus

- The forward equation (9.14a), by definition, are nothing more than the feedforward neural network description in (10.1). For example, the ith block of forward sub-equations (i.e. the ith layer equation) are

$$\boldsymbol{c}^i(\boldsymbol{u}, \mathbf{z}) = \boldsymbol{a}^i - \sigma\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right) = \mathbf{0}, \tag{10.3}$$

  the corresponding ith block of forward solution is $\boldsymbol{u}^i = \boldsymbol{a}^i$, and the ith block of parameter is $\mathbf{z}^i = \left[\mathbf{z}^i_{\boldsymbol{W}}, \mathbf{z}^i_{\boldsymbol{b}}\right]^T := \left[\boldsymbol{W}^i, \boldsymbol{b}^i\right]^T$. Clearly, the Jacobian $\nabla_{\boldsymbol{u}} \boldsymbol{c}$ is a lower block bi-diagonal matrix with identity blocks on the diagonal, and is thus invertible for all $\boldsymbol{u}$ and $\mathbf{z}$. Consequently, all results in Example 9.8 hold.
- To unfold the adjoint equation (9.14b), we note that the whole adjoint vector $\boldsymbol{v}$ is the concatenation of the adjoint sub-vector $\boldsymbol{v}^i$ corresponding to the ith layer equation in (10.1), for $i = 0, \ldots, \ell$. Thus, the ith adjoint equation corresponds to the derivative with respect to $\boldsymbol{u}^i = \boldsymbol{a}^i$ in (9.14b), and it reads

$$\begin{aligned} \boldsymbol{v}^{\ell} &= \boldsymbol{a}^{obs} - \boldsymbol{a}^{\ell}, \\ \boldsymbol{v}^{i} &= \left(\boldsymbol{W}^{i+1}\right)^{T} \left[\sigma'\left(\boldsymbol{W}^{i+1}\boldsymbol{a}^{i} + \boldsymbol{b}^{i+1}\right) \circ \boldsymbol{v}^{i+1}\right], \quad i = \ell - 1, \ldots, 0 \end{aligned} \tag{10.4}$$

where σ' is the derivative of σ, and $\circ$ denotes the component-wise multiplication of two vectors. Note that since σ acts componentwise (when its input is a vector), so is its derivative σ': in particular $\sigma'\left(\boldsymbol{W}^{i+1}\boldsymbol{a}^{i} + \boldsymbol{b}^{i+1}\right)$ is a vector in (10.4). Thus, (10.4) provides explicit expressions for the adjoint equations for a general fully connected DNN. Again, note that the adjoint equations are linear in terms of adjoint variables $\boldsymbol{v}^{i}$, $i = 0, \ldots, \ell$.

- To unfold the control equation (9.13) to explicitly see the derivative of the objective function with respect to the weights and biases, we take a block of control equations corresponding to sub-blocks of $\mathbf{z}^{i} = \left[\mathbf{z}^{i}_{\boldsymbol{W}}, \mathbf{z}^{i}_{\boldsymbol{b}}\right]^{T}$ in $\mathbf{z}$. For DNN, these derivatives are given as: for $i = 1, \ldots, \ell$,

$$\begin{aligned} \frac{\partial J}{\partial \boldsymbol{W}^{i}} &= -\left[\boldsymbol{v}^{\boldsymbol{i}} \circ \sigma'\left(\boldsymbol{W}^{i}\boldsymbol{a}^{i-1} + \boldsymbol{b}^{i}\right)\right]\left(\boldsymbol{a}^{i-1}\right)^{T}, \\ \frac{\partial J}{\partial \boldsymbol{b}^{i}} &= -\boldsymbol{v}^{i} \circ \sigma'\left(\boldsymbol{W}^{i}\boldsymbol{a}^{i-1} + \boldsymbol{b}^{i}\right), \end{aligned} \tag{10.5}$$

and the collection (in fact the concatenation) of these derivatives is the total reduced gradient ∇J.

The **backpropagation** nature of the network gradient is now clearly seen from the gradient expressions in (10.5) and the adjoint equations (10.4). Indeed, in (10.5) we need the ith adjoint state $\boldsymbol{v}^{i}$ in order to compute the gradients with respect to the weights and biases in the ith layer. The ith adjoint state $\boldsymbol{v}^{i}$ in turn depends on the $(i+1)$th adjoint state $\boldsymbol{v}^{i+1}$, which depends on the $(i+2)$th adjoint state $\boldsymbol{v}^{i+2}$, etc, all the way to the last adjoint state $\boldsymbol{v}^{\ell}$ corresponding to the network output layer. In other words, using the adjoint equations (10.4) we backpropagate to compute the adjoint solution from the output layer to the ith layer, and then compute the gradients using (10.5). From the backpropagation point of view, $\boldsymbol{v}^{i}$, $i = 1, \ldots, \ell$ are simply the temporary variables to help compute/write the chain rule in a succinct manner. *The adjoint approach, however, reveals their precise role as the adjoint solutions—also known as the Lagrangian multipliers—of the adjoint equations steming from the first order optimality condition using the reduced space approach in Remark 9.3.* Another important view point that we have exploited here is that the DNN training problem, from the adjoint point of view, is a constrained optimization problem with the forward pass as the forward equations. *The backpropagation is thus nothing more than a reduced space approach to compute the gradient using adjoint method.*

10.2 Hessian-vector product for deep neural networks

This section shows how to compute the Hessian-vector products exactly and efficiently. This knowledge allows us to deploy high-order optimization methods (such as Newton-type approach) without wary to train DNN. The materials that this section relies on are section 10.1 and section 20.1, and until then the readers can skip this section.

We now apply the development in section 20.1 to derive the full and Gauss-Newton Hessian-vector products for the optimization problem of training neural network with loss/objective/cost function (10.2) and equality constraints (10.3). Note that the gradients of weights and biases (10.5) is the total reduced gradient (20.5) when applied to the DNN training optimization problem. Similarly, the feedforward DNN (10.3) and the adjoint DNN (10.4) correspond to (20.6a) and (20.6b), respectively. Using the same notations, we define by $\widehat{(\cdot)}$ the total directional derivative of any quantity $(\cdot)$ with respect to $\boldsymbol{z}$ in a given direction $\hat{\boldsymbol{z}}$, where $\hat{\boldsymbol{z}} = \left[\widehat{\boldsymbol{W}^1}, \widehat{\boldsymbol{b}^1}, \ldots, \widehat{\boldsymbol{W}^\ell}, \widehat{\boldsymbol{b}^\ell}\right]^T$. In particular, the Hessian-vector product $\boldsymbol{H}_F \hat{\boldsymbol{z}}$ is precisely $\widehat{\nabla J}$: the directional derivative of the gradient in the given direction $\hat{\boldsymbol{z}}$. Thus, $\boldsymbol{H}_F \hat{\boldsymbol{z}}$ has the same number of components as ∇J does. Specifically, by the chain rules, the components of $\boldsymbol{H}_F \hat{\boldsymbol{z}}$ correspond to the components (10.5) of the gradient ∇J are given as: for $i = 1, \ldots, \ell$,

$$\widehat{\frac{\partial J}{\partial \boldsymbol{W}^i}} = -\left[\widehat{\boldsymbol{v}^i} \circ \sigma'\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right)\right]\left(\boldsymbol{a}^{i-1}\right)^T - \left[\boldsymbol{v}^i \circ \widehat{\sigma'}\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right)\right]\left(\boldsymbol{a}^{i-1}\right)^T - \left[\boldsymbol{v}^i \circ \sigma'\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right)\right]\left(\widehat{\boldsymbol{a}^{i-1}}\right)^T \quad (10.6a)$$

$$\widehat{\frac{\partial J}{\partial \boldsymbol{b}^i}} = -\widehat{\boldsymbol{v}^i} \circ \sigma'\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right) - \boldsymbol{v}^i \circ \widehat{\sigma'}\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right), \quad (10.6b)$$

where the j-component of the vector $\widehat{\sigma'}\left(\boldsymbol{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right)$ is given by

$$\sigma''\left(\boldsymbol{W}^i(j,:)\boldsymbol{a}^{i-1} + \boldsymbol{b}^i(j)\right)\widehat{\boldsymbol{W}^i}(j,:)\boldsymbol{a}^{i-1} + \sigma''\left(\boldsymbol{W}^i(j,:)\boldsymbol{a}^{i-1} + \boldsymbol{b}^i(j)\right)\widehat{\boldsymbol{b}^i}(j) + \sigma''\left(\boldsymbol{W}^i(j,:)\boldsymbol{a}^{i-1} + \boldsymbol{b}^i(j)\right)\boldsymbol{W}^i(j,:)\widehat{\boldsymbol{a}^{i-1}}, \quad (10.7)$$

where, recall our conventions, $\boldsymbol{W}^i(j,:)$ is the $j$the row of the weight matrix $\boldsymbol{W}^i$ and $\boldsymbol{b}^i(j)$ is the jth component of the bias vector $\boldsymbol{b}^i$.

Note that the components of the Hessian-vector product are not defined without computing the directional derivative of the states $\hat{\boldsymbol{u}} = \left[\widehat{\boldsymbol{a}^0}, \ldots, \widehat{\boldsymbol{a}^\ell}\right]^T$ and adjoints $\widehat{\boldsymbol{v}}$ first. They are solutions of the second order system (20.8). For DNN training, (20.8a) reduces to

$$\widehat{\boldsymbol{a}^i} - \widehat{\sigma}\left(\boldsymbol{W}^i\boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right) = \boldsymbol{0}, \quad i = 1, \ldots, \ell,$$

where, similar to (10.7), the j-component of the vector $\widehat{\sigma}\left(\boldsymbol{W}^i\boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right)$ is given by

$$\sigma'\left(\boldsymbol{W}^i(j,:)\boldsymbol{a}^{i-1} + \boldsymbol{b}^i(j)\right)\widehat{\boldsymbol{W}^i}(j,:)\boldsymbol{a}^{i-1} + \sigma'\left(\boldsymbol{W}^i(j,:)\boldsymbol{a}^{i-1} + \boldsymbol{b}^i(j)\right)\widehat{\boldsymbol{b}^i}(j) + \sigma'\left(\boldsymbol{W}^i(j,:)\boldsymbol{a}^{i-1} + \boldsymbol{b}^i(j)\right)\boldsymbol{W}^i(j,:)\widehat{\boldsymbol{a}^{i-1}},$$

Similarly, (20.8b) becomes

$$\begin{aligned}
\widehat{\boldsymbol{v}^\ell} &= -\widehat{\boldsymbol{a}^\ell}, \\
\widehat{\boldsymbol{v}^i} &= \left(\widehat{\boldsymbol{W}^{i+1}}\right)^T\left[\sigma'\left(\boldsymbol{W}^{i+1}\boldsymbol{a}^i + \boldsymbol{b}^{i+1}\right) \circ \boldsymbol{v}^{i+1}\right] \\
&\quad + \left(\boldsymbol{W}^{i+1}\right)^T\left[\widehat{\sigma'}\left(\boldsymbol{W}^{i+1}\boldsymbol{a}^i + \boldsymbol{b}^{i+1}\right) \circ \boldsymbol{v}^{i+1}\right] \\
&\quad + \left(\boldsymbol{W}^{i+1}\right)^T\left[\sigma'\left(\boldsymbol{W}^{i+1}\boldsymbol{a}^i + \boldsymbol{b}^{i+1}\right) \circ \widehat{\boldsymbol{v}^{i+1}}\right], \quad i = \ell - 1, \ldots, 0,
\end{aligned}$$

where $\widehat{\sigma'}\left(\boldsymbol{W}^{i+1}\boldsymbol{a}^i + \boldsymbol{b}^{i+1}\right)$ is given in (10.7) with $i+1$ in place of i.

For the Gauss-Newton Hessian-vector product, we simply remove second order derivatives terms to obtain

$$\begin{aligned}
\widehat{\frac{\partial J}{\partial \boldsymbol{W}^i}} &= -\left[\widehat{\boldsymbol{v}^i} \circ \sigma'\left(\boldsymbol{W}^i\boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right)\right]\left(\boldsymbol{a}^{i-1}\right)^T \\
\widehat{\frac{\partial J}{\partial \boldsymbol{b}^i}} &= -\widehat{\boldsymbol{v}^i} \circ \sigma'\left(\boldsymbol{W}^i\boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right),
\end{aligned}$$

where $\hat{\boldsymbol{v}}$, together with $\hat{\boldsymbol{u}} = \left[\widehat{\boldsymbol{a}^0}, \ldots, \widehat{\boldsymbol{a}^\ell}\right]^T$, is the solution of the simplified second order system (20.10). For the DNN training problem, we have

$$\widehat{\boldsymbol{a}^i} - \widehat{\sigma}\left(\boldsymbol{W}^i\boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right) = \boldsymbol{0}, \quad i = 1, \ldots, \ell,$$

and

$$\begin{aligned}
\widehat{\boldsymbol{v}^\ell} &= -\widehat{\boldsymbol{a}^\ell}, \\
\widehat{\boldsymbol{v}^i} &= \left(\boldsymbol{W}^{i+1}\right)^T\left[\sigma'\left(\boldsymbol{W}^{i+1}\boldsymbol{a}^i + \boldsymbol{b}^{i+1}\right) \circ \widehat{\boldsymbol{v}^{i+1}}\right], \quad i = \ell - 1, \ldots, 0.
\end{aligned}$$

Chapter 11
Solutions of linear least squares problems with adjoint

Abstract In Chapter 6 we have seen that the equation $\mathscr{A}u = f$ has a solution iff $f \in \mathsf{N}(\mathscr{A}^*)^\perp$ and when $\mathsf{N}(\mathscr{A}) = \{\theta\}$ the solution is unique. There are many practical situations in which a solution may not exist, and when it does there could be many of them. The question is how to choose a solution for the former and how to find a reasonable solution for the latter. Least squares provide a solution for these situations. Though it could be traced back as far as Gauss, Lengendre was the one who popularized least squares [135]. In this short chapter we will develop an abstract theory for linear least square using adjoint method. The keys that we rely on are a classical projection theorem and the closed range Theorem 5.2. The beauty here is that this approach allows us to obtain a general optimality condition for least squares that is valid for both finite and infinite dimensional settings without resorting to differentiation. The limitation is that this chapter is applicable to only linear least squares problems. We present several examples in finite dimensions to demonstrate the result. The application of the least squares theory in infinite dimensions will be discussed in Chapter 14 in the context of Tikhonov regularization approach. The extension to nonlinear problems will be discussed in Chapter 9, Chapter 10, Chapter 17, and Chapter 18.

We start with the classical projection theorem that holds for both finite and infinite dimensional settings.

Theorem 11.1 (Projection theorem). *Let $\mathcal{S}$ be a subspace of a pre-Hilbert*[1] *space $\mathbb{Y}$. Let $y \in \mathbb{Y}$. Then, $u \in \mathcal{S}$ is the unique minimizer of $\inf_{w \in \mathcal{S}} \|w - y\|_{\mathbb{Y}}$ iff $(y - u) \perp \mathcal{S}$. The existence of the minimizer u is guaranteed if $\mathbb{Y}$ is Hilbert and $\mathcal{S}$ is closed.*

Proof. We follow closely the proof by contradiction in [97, Theorem 2] and [109, Theorem 5.14.4]. Suppose there exists $v \in \mathcal{S}$ that is not orthogonal to $(y - u)$. We can assume that $(y - u, v)_{\mathbb{X}} = \varepsilon \neq 0$ and $\|v\|_{\mathbb{X}} = 1$. We have

[1] A pre-Hilbert space is an incomplete metric space with an inner product.

$$\|y-u-\varepsilon v\|_{\mathbb{Y}}^2 = \|y-u\|_{\mathbb{Y}}^2 + |\varepsilon|^2 - 2\varepsilon\bar{\varepsilon} = \|y-u\|_{\mathbb{Y}}^2 - |\varepsilon|^2 < \|y-u\|_{\mathbb{Y}}^2,$$

contradicting the fact that u is a minimizer. Conversely, let $(y-u) \perp \mathcal{S}$, by the Pythagorean identity (7.3), for any $v \in \mathcal{S}$ we have

$$\|y-v\|_{\mathbb{Y}}^2 = \|y-u+u-v\|_{\mathbb{Y}}^2 = \|y-u\|_{\mathbb{Y}}^2 + \|u-v\|_{\mathbb{Y}}^2 \geq \|y-u\|_{\mathbb{Y}}^2,$$

which shows that u is the unique minimizer. The existence proof is lengthy and hence is omitted.

The following corollary is also valid for both finite and infinite dimensions.

Corollary 11.1 (Linear least squares). *Let $\mathbb{X}, \mathbb{Y}$ be pre-Hilbert and $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be linear. Then, for any $y \in \mathbb{Y}$, $\tilde{x} \in \mathbb{X}$ is a minimizer of $\inf_{x\in\mathbb{X}} \|\mathscr{A}x - y\|_{\mathbb{Y}}$ iff*

$$\mathscr{A}^*\mathscr{A}\tilde{x} = \mathscr{A}^* y. \tag{11.1}$$

Furthermore, if $\mathscr{A}$ is injective then the minimizer $\tilde{x}$ is unique.

Proof. Note that $\inf_{x\in\mathbb{X}} \|\mathscr{A}x - y\|_{\mathbb{Y}}$ is equivalent to $\inf_{w\in\mathsf{R}(\mathscr{A})} \|w - y\|_{\mathbb{Y}}$. Applying Theorem 11.1 we know that $u \in \mathsf{R}(\mathscr{A})$ is a minimizer of $\inf_{w\in\mathsf{R}(\mathscr{A})} \|w - y\|_{\mathbb{Y}}$ iff $(y-u) \perp \mathsf{R}(\mathscr{A})$, that is, by the Closed Range Theorem 5.2,

$$(y-u) \in \mathsf{N}(\mathscr{A}^*) \Leftrightarrow \mathscr{A}^*(y-u) = \theta.$$

Since $u \in \mathsf{R}(\mathscr{A})$, there exists $\tilde{x} \in \mathbb{X}$ such that

$$\mathscr{A}^*(y - \mathscr{A}\tilde{x}) = \theta,$$

which concludes the first assertion.

For the second assertion, the uniqueness in Theorem 11.1 leads to the uniqueness of $\tilde{\boldsymbol{x}}$ when $\mathscr{A}$ is injective. Another way to see this is to note that $\mathsf{N}(\mathscr{A}^*\mathscr{A}) = \mathsf{N}(\mathscr{A})$ and thus the injectivity of $\mathscr{A}$ is equivalent to the injectivity of $\mathscr{A}^*\mathscr{A}$. The least squares solution in (11.1) is therefore unique.

The beauty here is that (11.1) is exactly the first order optimality condition that is typically obtained by requiring the derivative of $\|\mathscr{A}x - y\|_{\mathbb{Y}}$, with respective to x, to vanish (see the same result via derivative for linear least squares problem in Example 9.4). When $\dim(\mathbb{X}) < \infty$, then $\mathsf{R}(\mathscr{A})$ is finite dimension and hence closed in $\mathbb{Y}$. If, additionally, $\mathbb{Y}$ is Hilbert then the existence of $\tilde{x}$ is guaranteed by Theorem 11.1.

Example 11.1 (linear least squares in finite dimensions). Consider the operator $\mathscr{A}$ defined in Example 6.1 and we are interested in minimizing $\|\mathscr{A}x - \boldsymbol{y}\|_{\mathbb{R}^2}$ for some given $\boldsymbol{y} \in \mathbb{R}^2$. By Corollary 11.1, a minimizer $\tilde{x}$ must satisfy

$$\mathscr{A}^*\mathscr{A}\tilde{x} = \mathscr{A}^*\boldsymbol{y}.$$

Since $\mathscr{A}$ is not injective, there are multiple minimizers for this problem. This is consistent with the non-uniqueness in Example 6.1. Note that we can reformulate the operator form $\|\mathscr{A}x - \boldsymbol{y}\|_{\mathbb{R}^2}$ by an equivalent matrix representation form. Indeed, let $\boldsymbol{A}$ be the matrix representation of $\mathscr{A}$ with respect to an orthonormal basis in $\mathbb{U} = \text{Span}\{1,\ x,\ x^2\}$ and the canonical basis of $\mathbb{R}^2$. Let $\boldsymbol{x}$ be the coordinate vector of x in the same orthonormal basis of $\mathbb{U} = \text{Span}\{1,\ x,\ x^2\}$, we have $\|\mathscr{A}x - \boldsymbol{y}\|_{\mathbb{R}^2} = \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^2}$. Again, $\boldsymbol{A}$ is not injective and a solution is not unique.

Example 11.2 (linear least squares with matrices). Consider $\boldsymbol{A} : \mathbb{F}^n \to \mathbb{F}^m$ and $\boldsymbol{y} \in \mathbb{F}^m$. Applying Corollary 11.1 we have that there exists $\tilde{\boldsymbol{x}} \in \mathbb{F}^n$ minimizing $\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{F}^m}$ and

$$\boldsymbol{A}^*\boldsymbol{A}\tilde{\boldsymbol{x}} = \boldsymbol{A}^*\boldsymbol{y}.$$

When $\boldsymbol{A}$ is injective or equivalently full column rank, we have the uniqueness of $\tilde{\boldsymbol{x}}$ again by Corollary 11.1. Another way to see this is that in this case, $\boldsymbol{A}^*\boldsymbol{A}$ is an invertible matrix which implies the uniqueness of the minimizer $\tilde{\boldsymbol{x}}$.

Example 11.3 (Linear regression). Suppose that we are given a data set with N data points $\left\{\boldsymbol{z}^i, t^i\right\}_{i=1}^N$, where $\boldsymbol{z}^i \in \mathbb{R}^d$ and $y^i \in \mathbb{R}$. We wish to fit the data with a hyperplane $\hat{t}\left(\boldsymbol{z}\right) := \sum_{j=1}^{d} \boldsymbol{\theta}_j \boldsymbol{z}_j + \gamma = \boldsymbol{\theta}^T\boldsymbol{z} + \gamma$, where $\boldsymbol{\theta}$ and γ are to be determined. The best unknowns, in the least squares sense, are the ones that solve the following optimization problem

$$\min_{\boldsymbol{\theta},\gamma} \sum_{i=1}^{N} \left[\hat{t}\left(\boldsymbol{z}^i\right) - t^i\right]^2 = \min_{\boldsymbol{\theta},\gamma} \sum_{i=1}^{N} \left(\sum_{j=1}^{d} \boldsymbol{\theta}_j \boldsymbol{z}_j^i + \gamma - t^i\right)^2 \tag{11.2}$$

In order to cast the linear regression problem (11.2) into the form discussed above, let us concatenate $\boldsymbol{\theta}$ and γ into a single unknown vector $\boldsymbol{x} := \left[\gamma, \boldsymbol{\theta}^T\right] \in \mathbb{R}^{d+1}$, $\boldsymbol{y} := \left[t^1, \ldots, t^N\right]$, and define the feature matrix

$$\boldsymbol{A} := \begin{bmatrix} 1 & \boldsymbol{z}_1^1 & \ldots & \boldsymbol{z}_d^1 \\ 1 & \boldsymbol{z}_1^2 & \ldots & \boldsymbol{z}_d^2 \\ \vdots & & & \\ 1 & \boldsymbol{z}_1^N & \ldots & \boldsymbol{z}_d^N \end{bmatrix}.$$

The linear regression problem (11.2) becomes $\min_{\boldsymbol{x}\in\mathbb{R}^{d+1}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^N}^2$, which is equivalent to $\min_{\boldsymbol{x}\in\mathbb{R}^{d+1}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^N}$. As we have shown, an optimal solution $\boldsymbol{x}$ is given by

$$\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{x} = \boldsymbol{A}^T\boldsymbol{y},$$

since $\boldsymbol{A}^* = \boldsymbol{A}^T$. If, in addition, the feature matrix $\boldsymbol{A}$ is full column rank, then the optimal solution $\boldsymbol{x} = \left(\boldsymbol{A}^T\boldsymbol{A}\right)^{-1}\boldsymbol{A}^T\boldsymbol{y}$ is unique. We can further show that this is, in fact, the best unbiased linear estimator (known as the Gauss-Markov theorem [79]), but since our focus here is the adjoint, we omit this detail.

Example 11.4 (Linear regression with nonlinear features). In Example 11.3, we consider ordinary least squares problems with linear feature. In this example, we consider a more general setting with nonlinear features. The idea is to first transform the original variable $\boldsymbol{z}$ to a feature variable via a transformation: $\boldsymbol{\Phi} : \mathbb{R}^d \to \mathbb{R}^m$. Then, the regression is performed in the feature space. In particular, we fit the data with a hyperplane $\hat{t}\left(\boldsymbol{z}\right) := \sum_{j=1}^{m} \boldsymbol{\theta}_j \boldsymbol{\Phi}_j\left(\boldsymbol{z}\right) + \gamma$. The linear regression problem is still of the same form $\min_{\boldsymbol{x}\in\mathbb{R}^{m+1}} \|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^N}$, but the feature matrix now reads

$$\boldsymbol{A} := \begin{bmatrix} 1 & \boldsymbol{\Phi}^T\left(\boldsymbol{z}^1\right)] \\ 1 & \boldsymbol{\Phi}^T\left(\boldsymbol{z}^2\right)] \\ \vdots & \\ 1 & \boldsymbol{\Phi}^T\left(\boldsymbol{z}^N\right)] \end{bmatrix}.$$

Example 11.5 (Least squares with constraints). We consider the settin in Example 6.1 in which $\boldsymbol{A} : \mathbb{R}^2 \to \mathbb{R}^2$ is given as

$$\boldsymbol{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}.$$

We have shown that there are many solutions to the equation

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f},$$

for any $\boldsymbol{f} \in \mathbb{R}^2$. We are interested in a solution with minimum norm, i.e.,

$$\min_{\boldsymbol{u}\in\mathbb{R}^2} \|\boldsymbol{u}\|_{\mathbb{R}^2}^2$$

subject to

$$\boldsymbol{A}\boldsymbol{u} = \boldsymbol{f}.$$

Part III
Adjoint operators in infinite dimensional Hilbert spaces

Use the template *part.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page in the Springer layout.

Chapter 12
Notations and conventions

Abstract
In this chapter, we shall consider linear mappings between infinite dimensionsal Hilbert spaces. Of particular interest are densely defined differential operators.

12.1 Adjoints of densely defined linear operators

We start with the definition of dense subsets based on the closure that we have defined in Definition 5.8.

Definition 12.1 (Dense subset). $\mathcal{S}$ is called a dense subset of $\mathbb{U}$ if its closure $\overline{\mathcal{S}}$ coincides with $\mathbb{U}$. In other words, for any $u \in \mathbb{U}$, there exists a sequence $\{u^n\}_{n=1}^{\infty} \subset \mathcal{S}$ converges to u under the $\mathbb{U}$-norm (i.e., $u^n \xrightarrow{\|\cdot\|_{\mathbb{U}}} u$).

An important application of the density for a densely defined continuous linear operator is the following (continuous) extension principle [82, 143]. Definition and some properties of continuous linear operators can be found in Chapter 5. In this chapter, we will invoke this extension principle to show the existence of adjoint operators (and it will be used in Chapter 13 to characterize $\mathbb{H}^{-m}$ spaces.).

Lemma 12.1 (Continuous extension principle). *Let $\mathbb{U}$ and $\mathbb{Y}$ be Banach spaces*[1] *over $\mathbb{F}$ (real or complex). Assume*

- *$\mathscr{A} : \mathsf{D}(\mathscr{A}) \subset \mathbb{U} \to \mathbb{V}$ is continuous on $\mathsf{D}(\mathscr{A})$, i.e., there exists a positive constant c such that*

$$\|\mathscr{A}u\|_{\mathbb{V}} \leq c\,\|u\|_{\mathbb{U}}\,, \quad \forall u \in \mathsf{D}(\mathscr{A})\,, \text{ and}$$

[1] A Banach space $\mathbb{X}$ is a normed vector space in which any Cauchy sequences converges to a well-defined limit residing in $\mathbb{X}$.

- $\mathsf{D}(\mathscr{A})$ *is dense in* $\mathbb{U}$.

Then there is a unique extension of $\mathscr{A}$*, again denoted by* $\mathscr{A}$*, such that* $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ *is a continuous linear operator.*

Proof. A proof of this standard result can be found in [143, Proposition 1 in section 3.6 of chapter 3].

For any operator $\mathscr{A} : \mathbb{U} \to \mathbb{V}$ considered in this chapter, we assume that its domain $\mathsf{D}(\mathscr{A})$ is dense in $\mathbb{U}$. In this case, we say that $\mathscr{A}$ is densely defined. A sufficient general definition of the adjoint operator of a densely defined operator that could embrace a variety of problems is the following (see, e.g., [132]).

Definition 12.2 (Adjoint operators). Let $\mathscr{A} \in \mathcal{L}(\mathbb{U}, \mathbb{V})$ be densely defined. $\mathscr{A}^*$ is called the adjoint of $\mathscr{A}$

$$(\mathscr{A}u, v)_{\mathbb{V}} = (u, \mathscr{A}^* v)_{\mathbb{U}} \quad \forall u \in \mathsf{D}(\mathscr{A}) \text{ and } \forall v \in \mathsf{D}(\mathscr{A}^*), \tag{12.1}$$

where

$$\mathsf{D}(\mathscr{A}^*) := \{v : \text{ the map } u \mapsto (\mathscr{A}u, v)_{\mathbb{V}} \text{ is continuous on } \mathbb{U}\}.$$

Let us provide some insights into the adjoint Definition 12.2. First, Definition 12.2 reduces to Definition 5.6 when $\mathscr{A}$ is continuous and is defined on the whole space $\mathbb{U}$. Second, $\mathsf{D}(\mathscr{A}^*)$ is a vector space from the assumption that $\mathsf{D}(\mathscr{A})$ is a vector space. Third, the existence of such an adjoint is clear by the definition of its domain. Indeed, for each $v \in \mathsf{D}(\mathscr{A}^*)$, $u \mapsto (\mathscr{A}u, v)_{\mathbb{V}}$ is continuous in $u \in \mathsf{D}(\mathscr{A})$. Since $\mathsf{D}(\mathscr{A})$ is dense in $\mathbb{U}$, by Lemma 12.1 there is a unique continuous extension of $(\mathscr{A}\cdot, v)_{\mathbb{V}}$ over the whole $\mathbb{U}$, again denoted by $(\mathscr{A}\cdot, v)_{\mathbb{V}}$. By the Riesz representation Theorem 5.1, there is a linear functional $\ell \in \mathbb{U}^*$ such that (see Remark 5.4)

$$(u, \ell)_{\mathbb{U}} = (\mathscr{A}u, v)_{\mathbb{V}}.$$

Since the right-hand side is (anti) linear in v, ℓ is a linear function in v. That is, the mapping from $v \in \mathbb{V}$ to $\ell \in \mathbb{U}$ is linear and let us call it $\mathscr{A}^*$ such that

$$\ell = \mathscr{A}^* v,$$

and thus we have

$$(u, \mathscr{A}^* v)_{\mathbb{U}} = (u, \ell)_{\mathbb{U}} = (\mathscr{A}u, v)_{\mathbb{V}},$$

which is exactly (12.1). Our argument shows that the adjoint operator $\mathscr{A}^*$ defined above is well-defined by the definition of its domain. This is a common way to define an operator (see many examples below). *If we change the*

domain, we change the operator. In other words, an operator is not defined until its domain is specified.

Example 12.1. Let us define the space of square integrable functions on $(0,1)$ as

$$\mathbb{L}^2(0,1) := \left\{ f : (0,1) \to \mathbb{R} \text{ such that } \int_0^1 |f(x)|^2 \, dx < \infty \right\}.$$

with the inner product

$$(f,g)_{\mathbb{L}^2(0,1)} = \int_0^1 f(x) g(x) \, dx, \quad \forall f, g \in \mathbb{L}^2(0,1), \tag{12.2}$$

and the induced $\mathbb{L}^2$-norm $\|u\|_{\mathbb{L}^2(0,1)} = \sqrt{(u,u)_{\mathbb{L}^2(0,1)}}$. We now define the action of a linear map $\mathscr{A}$ on a function $u \in \mathbb{L}^2(0,1)$ as the second order differentiation

$$\mathscr{A} u := u'' := \frac{d^2 u}{dx^2}$$

with the domain

$$\mathsf{D}(\mathscr{A}) := \left\{ u \in \mathcal{C}^2[0,1] : u(0) = u'(0) = 0 \right\},$$

where $\mathcal{C}^2[0,1]$ is the space of continuously twice differentiable functions. One can show that $\mathsf{D}(\mathscr{A})$ is dense in $\mathbb{L}^2(0,1)$, and thus $\mathscr{A} : \mathsf{D}(\mathscr{A}) \subset \mathbb{L}^2(0,1) \to \mathbb{L}^2(0,1)$ is densely defined. We are interested in finding the adjoint operator $\mathscr{A}^*$ as a linear map from $\mathbb{L}^2(0,1)$ to $\mathbb{L}^2(0,1)$ using Definition 12.2. For any $v \in \mathcal{C}^2[0,1]$, by integrating by parts twice, we have

$$(\mathscr{A} u, v)_{\mathbb{L}^2(0,1)} = (u, v'')_{\mathbb{L}^2(0,1)} + u' v|_0^1 - u v'|_0^1.$$

The boundary terms vanish and $(\mathscr{A} u, v)_{\mathbb{L}^2(0,1)}$ is continuous in u with respect to the $\mathbb{L}^2$-norm, if we define the domain of the adjoint operator $\mathscr{A}^*$ to be

$$\mathsf{D}(\mathscr{A}^*) := \left\{ v \in \mathcal{C}^2[0,1] : v(1) = v'(1) = 0 \right\}.$$

We conclude that the adjoint operator is also the same second order differentiation but with different domain.

Example 12.2. Consider the same setting as in Example 12.1, but the domain of the second order differentiable operator $\mathscr{A}$ is now

$$\mathsf{D}(\mathscr{A}) := \left\{ u \in \mathcal{C}^2[0,1] : u(0) = u(1) = 0 \right\}.$$

Following the same procedure, we can easily see that not only $\mathscr{A}^*$ is the same second order differentiation but also it has the same domain (see Problem 12.2). For such cases, we call $\mathscr{A}^*$ self-adjoint (see Definition 7.2).

We next discuss the uniqueness of $\mathscr{A}^*$. Assume that there are two adjoint operators $\mathscr{A}^*$ and $\widehat{\mathscr{A}^*}$. From Definition 12.2, clearly both $\mathscr{A}^*$ and $\widehat{\mathscr{A}^*}$ have the same domain $\mathsf{D}(\mathscr{A}^*)$. For any $v \in \mathsf{D}(\mathscr{A}^*)$, as discuss above on the existence of the adjoint operator $\mathscr{A}^*$, we have

$$\left(u, \mathscr{A}^* v - \widehat{\mathscr{A}^*} v\right)_{\mathbb{U}} = 0, \quad \forall u \in \mathbb{U},$$

which yields

$$\left(\mathscr{A}^* - \widehat{\mathscr{A}^*}\right) v = \theta, \quad \forall v \in \mathsf{D}(\mathscr{A}^*),$$

and thus $\mathscr{A}^* = \mathscr{A}$ on $\mathsf{D}(\mathscr{A}^*)$.

12.2 Adjoints of classical differential operators

In this section, we consider several examples of $\mathscr{A}$ as classical differential operators and determine the corresponding adjoints.

Example 12.3 (Divergent operator). Let Ω be an open and bounded subset of $\mathbb{R}^n$ with n as a natural number. The space of square-integrable vector-valued functions $\left[\mathbb{L}^2(\Omega)\right]^n$ is defined as (see Definition 13.7 for general $\mathbb{L}^p$ spaces)

$$\left[\mathbb{L}^2(\Omega)\right]^n := \left\{ \boldsymbol{f} : \Omega \to \mathbb{R} \text{ such that } \int_\Omega \|\boldsymbol{f}(\boldsymbol{x})\|_{\mathbb{R}^n}^2 \, d\boldsymbol{x} < \infty \right\}.$$

with the inner product

$$(\boldsymbol{f}, \boldsymbol{g})_{[\mathbb{L}^2(\Omega)]^n} = \int_\Omega \boldsymbol{f}(\boldsymbol{x}) \boldsymbol{g}(\boldsymbol{x}) \, d\boldsymbol{x}, \quad \forall \boldsymbol{f}, \boldsymbol{g} \in \left[\mathbb{L}^2(\Omega)\right]^n, \tag{12.3}$$

and the induced $\mathbb{L}^2$-norm $\|\boldsymbol{u}\|_{[\mathbb{L}^2(\Omega)]^n} = \sqrt{(\boldsymbol{u}, \boldsymbol{u})_{[\mathbb{L}^2(\Omega)]^n}}$. Note that when $n = 1$ the above definition reduces to scalar-valued functions. Again, we use normal and boldface letters for scalar-valued and vector-valued functions, respectively. Consider the classical gradient operator $\mathscr{A} : \mathbb{L}^2(\Omega) \to \left[\mathbb{L}^2(\Omega)\right]^n$

$$\mathscr{A} u := \nabla u,$$

with the domain

$$\mathsf{D}(\mathscr{A}) := \left\{ u \in \mathcal{C}^1\left(\overline{\Omega}\right) : u(\boldsymbol{x}) = 0 \text{ on } \partial\Omega \right\},$$

where the $\overline{\Omega}$ denote the closure of Ω in $\mathbb{R}^n$ (see Definition 5.8), and $\partial\Omega$ denotes the boundary of Ω. For any $\boldsymbol{v} \in \left[\mathcal{C}^1\left(\overline{\Omega}\right)\right]^n$, integrating by parts gives

$$(\mathscr{A}u, \boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n} = (\nabla u, \boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n} = (u, -\nabla \cdot \boldsymbol{v})_{\mathbb{L}^2(\Omega)} \leq \|u\|_{\mathbb{L}^2(\Omega)} \|\nabla \cdot \boldsymbol{v}\|_{\mathbb{L}^2(\Omega)},$$

where we have used the Cauchy-Schwarz inequality in the last inequality. It follows that $(\mathscr{A}u, \boldsymbol{v})_{\mathbb{L}^2(\Omega)}$ is continuous with respect to u in the $\mathbb{L}^2$-norm. Thus, by Definition 12.2, the adjoint of the classical gradient operator with the above domain is the negative divergence operator with $\mathcal{C}^1\left(\overline{\Omega}\right)$ as the domain.[2]

On the other hand, if we enlarge the domain of $\mathscr{A}$ to be $\mathcal{C}^1\left(\overline{\Omega}\right)$, then the domain of the adjoint $\mathscr{A}^*$ becomes

$$\mathsf{D}\left(\mathscr{A}^*\right) := \left\{\boldsymbol{v} \in \mathcal{C}^1\left(\overline{\Omega}\right) : \boldsymbol{v} \cdot \boldsymbol{n} = 0 \text{ on } \partial\Omega\right\},$$

which is smaller. Here, $\boldsymbol{n}$ denotes the outward unit normal vector on $\partial\Omega$. In this case, the adjoint $\mathscr{A}^*$ is still the negative divergence, but with smaller domain.

Remark 12.1. Since the adjoint of the adjoint is the original operator, we can see that the adjoint of the divergence operator with a specified domain is the negative gradient operator with an appropriate domain. In a similar fashion, one can easily show that the adjoint of $\mathscr{A} = \nabla\times$ is itself (with possibly a different domain of course depending on the domain of $\mathscr{A}$): see Problem 12.4.

Remark 12.2. Example 12.3 presents an important fact: the larger the $\mathsf{D}\left(\mathscr{A}\right)$ is, the smaller the $\mathsf{D}\left(\mathscr{A}^*\right)$, and vice versa. This will be the guiding principle for Chapter 13 in which we construct generalized derivatives of functions that do not have classical derivatives by enlarging the domain of the adjoint $\mathscr{A}^*$ while simultaneously reducing the domain of the original operator $\mathscr{A}$.

Note that we can define the domain of a differential operator with boundary values specified on a portion of the boundary $\partial\Omega$ (see Example 12.5. In that case, the domain of the adjoint typically has boundary values specified on the rest of the boundary. The key is to make sure that boundary terms that arise in the integration by parts vanish, all the volume and boundary integrals make sense, and $(\mathscr{A}u, v)_{\mathbb{U}}$ is continuous in u with respect to the $\mathbb{U}$-norm. We now carry out these steps for three important examples of partial differential equations.

Example 12.4 (An elliptic differential operator). With $\mathbb{L}^2(\Omega)$ defined as in Example 12.3, consider the following parametrized linear operator $\mathscr{A} : \mathsf{D}\left(\mathscr{A}\right) \subset \mathbb{L}^2(\Omega) \to \mathbb{L}^2(\Omega)$ as

$$\mathscr{A}u = -\nabla \cdot \left(e^z \nabla u\right) \text{ in } \Omega,$$

[2] Note that the adjoint operator still obeys Definition 12.2 with the possibly bigger domain as long as $\mathsf{D}\left(\mathscr{A}^*\right)$ is such that $\nabla \cdot \boldsymbol{v} \in \mathbb{L}^2(\Omega)$, and hence $(\mathscr{A}u, \boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n}$ is continuous with respect to u in the $\mathbb{L}^2$-norm, but we omit the details here.

where $z \in \mathcal{C}^1(\Omega) \subset \mathbb{L}^2(\Omega)$ is a distributed parameter over Ω with sufficiently smooth boundary $\partial\Omega$. We choose the domain of $\mathscr{A}$ to be

$$\mathsf{D}(\mathscr{A}) := \left\{u \in \mathcal{C}^2\left(\overline{\Omega}\right) : u(\boldsymbol{x}) = 0 \text{ on } \partial\Omega\right\}.$$

One can show the standard result [3, 102] that $\mathsf{D}(\mathscr{A})$ is dense in $\mathbb{L}^2(\Omega)$, but this is out of the scope of the book. For each $v \in \mathcal{C}^2\left(\overline{\Omega}\right)$, integrating by parts twice yields

$$\begin{aligned}(\mathscr{A}u, v)_{\mathbb{L}^2(\Omega)} &= -\left(\nabla \cdot \left(e^z \nabla u\right), v\right)_{\mathbb{L}^2(\Omega)} \\ &= -\left(\nabla \cdot \left(e^z \nabla v\right), u\right)_{\mathbb{L}^2(\Omega)} - \int_{\partial\Omega} e^{z(\boldsymbol{x})} \nabla u(\boldsymbol{x}) \cdot \boldsymbol{n}\, v(\boldsymbol{x})\, d\boldsymbol{x}.\end{aligned}$$

If we restrict $v \in \mathsf{D}(\mathscr{A})$, we have

$$(\mathscr{A}u, v)_{\mathbb{L}^2(\Omega)} = -\left(\nabla \cdot \left(e^z \nabla v\right), u\right)_{\mathbb{L}^2(\Omega)} \leq \|u\|_{\mathbb{L}^2(\Omega)} \left\|\nabla \cdot \left(e^z \nabla v\right)\right\|_{\mathbb{L}^2(\Omega)},$$

and by Definition 12.2, $\mathsf{D}(\mathscr{A}^*) = \mathsf{D}(\mathscr{A})$, and $\mathscr{A}^* = \mathscr{A}$. Thus, $\mathscr{A}$ is self-adjoint.

Example 12.5 (A hyperbolic differential operator). With $\mathbb{L}^2(\Omega)$ defined as in Example 12.3, consider the following parametrized linear operator $\mathscr{A} : \mathsf{D}(\mathscr{A}) \subset \mathbb{L}^2(\Omega) \to \mathbb{L}^2(\Omega)$ as

$$\mathscr{A}u = \boldsymbol{\beta} \cdot \nabla u + \lambda u \text{ in } \Omega,$$

where $\boldsymbol{\beta} \in \left[\mathcal{C}^1(\Omega, \mathbb{R})\right]^n$ and $\nabla \cdot \boldsymbol{\beta} = 0$, $\lambda > 0$, $\boldsymbol{n}$ is the unit outward normal vector of the boundary $\partial\Omega$, and $\partial\Omega_{\text{in}} := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{\beta} \cdot \boldsymbol{n} < 0\}$ is the inflow boundary. For $\mathscr{A}$ to make sense in the classical sense, we choose its domain to be

$$\mathsf{D}(\mathscr{A}) := \left\{u \in \mathcal{C}^1\left(\overline{\Omega}\right) : u(\boldsymbol{x}) = 0 \text{ in } \partial\Omega_{\text{in}}\right\}.$$

One can show that $\mathsf{D}(\mathscr{A})$ is dense in $\mathbb{L}^2(\Omega)$, but this is not of interest here (see Example 12.9 for more information). For each $v \in \mathcal{C}^1\left(\overline{\Omega}\right)$, integrating by parts once gives

$$\begin{aligned}(\mathscr{A}u, \boldsymbol{v})_{\mathbb{L}^2(\Omega)} &= (\boldsymbol{\beta} \cdot \nabla u + \lambda u, v)_{\mathbb{L}^2(\Omega)} \\ &= (u, -\boldsymbol{\beta} \cdot \nabla v + \lambda v)_{\mathbb{L}^2(\Omega)} + \int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta} \cdot \boldsymbol{n} u\, v\, d\boldsymbol{x},\end{aligned}$$

where the outflow boundary is defined as $\partial\Omega_{\text{in}} := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{\beta} \cdot \boldsymbol{n} > 0\}$. Thus, if we restrict v in

$$\mathcal{S} := \left\{v \in \mathcal{C}^1\left(\overline{\Omega}\right) : v(\boldsymbol{x}) = 0 \text{ in } \partial\Omega_{\text{out}}\right\},$$

so that the boundary term vanishes, then we have

$$(\mathscr{A}u, \boldsymbol{v})_{\mathbb{L}^2(\Omega)} = (u, -\boldsymbol{\beta}\cdot\nabla v + \lambda v)_{\mathbb{L}^2(\Omega)} \leq \|u\|_{\mathbb{L}^2(\Omega)} \|-\boldsymbol{\beta}\cdot\nabla v + \lambda v\|_{\mathbb{L}^2(\Omega)},$$

which means $(\mathscr{A}u, \boldsymbol{v})_{\mathbb{L}^2(\Omega)}$ is continuous in u with respect to the $\mathbb{L}^2$-norm. By Definition 12.2, the adjoint operator $\mathscr{A}^* : \mathsf{D}(\mathscr{A}^*) \subset \mathbb{L}^2(\Omega) \to \mathbb{L}^2(\Omega)$ is

$$\mathscr{A}^* v = -\boldsymbol{\beta}\cdot\nabla v + \lambda v,$$

for any $v \in \mathsf{D}(\mathscr{A}^*) = \mathcal{S}$.

Example 12.6 (Friedrichs differential operators). We are interested in Friedrichs system that embraces a large class of elliptic, parabolic, hyperbolic, and mixed-type PDEs operators [58]:

$$\mathscr{A}\boldsymbol{u} := \sum_{k=1}^{n} \boldsymbol{A}_k \partial_k \boldsymbol{u} + \boldsymbol{C}\boldsymbol{u} \text{ in } \Omega,$$

where d is the spatial dimension, $\boldsymbol{u}$ the vector-valued unknown solution in $\mathbb{R}^m$, and Ω is an open and bounded subset of $\mathbb{R}^n$ with sufficient regular boundary $\partial\Omega$. The matrices $\boldsymbol{A}_k$ and $\boldsymbol{C}$ are assumed to be continuous across Ω. Here, ∂_k is understood as the kth partial derivative. We start with the standard assumptions (see, e.g., [58, 52, 78, 54]):

$$\begin{aligned}
&\boldsymbol{C} \in \left[\mathcal{C}\left(\overline{\Omega}\right)\right]^{m,m},\\
&\boldsymbol{A}_k \in \left[\mathcal{C}\left(\overline{\Omega}\right)\right]^{m,m}, \quad k = 1,\ldots,n, \quad \text{and } \sum_{k=1}^{n} \partial_k \boldsymbol{A}_k \in \left[\mathcal{C}\left(\overline{\Omega}\right)\right]^{m,m},\\
&\boldsymbol{A}_k = \left(\boldsymbol{A}_k\right)^T \text{ in } \Omega, \quad k = 1,\ldots,n,,\\
&\boldsymbol{C} + \boldsymbol{C}^T + \sum_{k=1}^{n} \partial_k \boldsymbol{A}_k \geq 2\alpha_0 I \text{ in } \Omega,
\end{aligned}$$

where $\alpha_0 > 0$ is some coercivity constant.

Next, we specify an abstract boundary condition for $\mathscr{A}$ [52]. Let $\boldsymbol{D} := \sum_{k=1}^{n} \boldsymbol{A}_k \boldsymbol{n}_k$, where $\boldsymbol{n}_k$ is the kth component of the unit outward normal vector $\boldsymbol{n}$ on $\partial\Omega$, and assume there exists a matrix $\boldsymbol{M} \in \left[\mathcal{C}\left(\overline{\partial\Omega}\right)\right]^{m,m}$ such that

$$\begin{aligned}
\boldsymbol{M} + \boldsymbol{M}^T &\geq 0 \text{ on } \partial\Omega,\\
(\boldsymbol{D} - \boldsymbol{M})\,\boldsymbol{u} &= \boldsymbol{0} \text{ on } \partial\Omega,\\
\mathsf{N}(\boldsymbol{D} - \boldsymbol{M}) + \mathsf{N}(\boldsymbol{D} + \boldsymbol{M}) &= \mathbb{R}^m, \text{ on } \partial\Omega.
\end{aligned} \tag{12.4}$$

We define the domain of $\mathscr{A}$ as

$$\mathsf{D}(\mathscr{A}) := \left\{\boldsymbol{u} \in \left[\mathcal{C}^1\left(\overline{\Omega}\right)\right]^n : (\boldsymbol{D} - \boldsymbol{M})\,\boldsymbol{u} = \boldsymbol{0} \text{ on } \partial\Omega\right\},$$

which can be shown to be dense in $\left[\mathbb{L}^2(\Omega)\right]^n$ [54, 102, 3]. Taking $\boldsymbol{v} \in \left[\mathcal{C}^1\left(\overline{\Omega}\right)\right]^n$ and integrating by parts once give

$$(\mathscr{A}\boldsymbol{u}, \boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n} = (\boldsymbol{u}, \mathscr{B}\boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n} + (\boldsymbol{u}, \boldsymbol{D}\boldsymbol{v})_{\mathbb{L}^2(\partial\Omega)},$$

for all $u \in \mathsf{D}(\mathscr{A})$, where

$$\mathscr{B}\boldsymbol{v} := -\sum_{k=1}^{n} \partial_k \left(\boldsymbol{A}_k \boldsymbol{v}\right) + \boldsymbol{C}^T \text{ in } \Omega.$$

Now invoking the boundary conditions (12.4), it is easy to see that the boundary term vanishes, i.e.,

$$(\boldsymbol{u}, \boldsymbol{D}\boldsymbol{v})_{\mathbb{L}^2(\partial\Omega)}$$

when $\boldsymbol{v}$ satisfies $\boldsymbol{D}\boldsymbol{v} \in \mathsf{N}(\boldsymbol{D} - \boldsymbol{M})^{\perp}$. Thus, if we choose the domain of the adjoint operator $\mathscr{A}^*$ to be

$$\mathsf{D}(\mathscr{A}^*) = \left\{\boldsymbol{v} \in \left[\mathcal{C}^1\left(\overline{\Omega}\right)\right]^n : \boldsymbol{D}\boldsymbol{v} \in [\mathsf{N}(\boldsymbol{D} - \boldsymbol{M})]^{\perp}\right\},$$

the by definition, $\mathscr{B}$ is the adjoint $\mathscr{A}^*$ since

$$(\mathscr{A}\boldsymbol{u}, \boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n} = (\boldsymbol{u}, \mathscr{B}\boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n} \leq \|\boldsymbol{u}\|_{[\mathbb{L}^2(\Omega)]^n} \|\mathscr{B}\boldsymbol{v}\|_{[\mathbb{L}^2(\Omega)]^n},$$

that is, $(\mathscr{A}\boldsymbol{u}, \boldsymbol{v})_{[\mathbb{L}^2(\Omega)]^n}$ is continuous in $\boldsymbol{u}$ with respect to the $\left[\mathbb{L}^2(\Omega)\right]^n$-norm.

In this section, we consider several examples of adjoint operators of linear differential operators. By choosing the domain of a differential operator $\mathscr{A}$ as a closed subset of the graph space $\mathcal{G}_{\mathscr{A}}$, $\mathscr{A}$ is automatically continuous in the graph norm. The adjoint is thus continuous and has the same norm (see Proposition 5.3).

Come back to this after the Green chapter. We should only consider closed differential operators here by choosing the domain as $\mathcal{C}^n$, where n is the order of the differential operator. Then show that all the differential operators are closed in the $\mathbb{L}^2$-setting. Only move to the graph space to have continuous operators in the BNB chapter. So essentially we move all the following examples to the BNB chapter and only consider its closed versions in this chapter.

Example 12.7. We consider an elliptic differential operator in n dimensions over an open and bounded domain $\Omega \subset \mathbb{R}^n$. In this case, we define $\mathbb{L}^2-$inner product of two functions $u(\boldsymbol{x}), v(\boldsymbol{x})$ in $\mathbb{L}^2(\Omega)$ over $\mathbb{R}$ as

$$(u, v)_{L^2(\Omega)} = \int_{\Omega} u(\boldsymbol{x}) v(\boldsymbol{x}) d\Omega. \tag{12.5}$$

Consider the following parametrized linear operator $\mathscr{A} : \mathsf{D}(\mathscr{A}) \subset \mathbb{L}^2(\Omega) \to \mathbb{L}^2(\Omega)$ as

$$\mathscr{A}u = \begin{cases} -\nabla \cdot (e^z \nabla u) & \text{in } \Omega, \\ u = 0 & \text{on } \partial\Omega, \end{cases}$$

where $z \in \mathcal{C}^1(\Omega) \subset \mathbb{L}^2(\Omega)$ is a distributed parameter over Ω with sufficiently smooth boundary $\partial\Omega$. We define the graph space $\mathcal{G}_{\mathscr{A}}$ as

$$\mathcal{G}_{\mathscr{A}} := \left\{u \in \mathbb{H}^1(\Omega) : \mathscr{A}u \in \mathbb{L}^2(\Omega)\right\} = \left\{u \in \mathbb{H}^1(\Omega) : -\nabla \cdot (e^z \nabla u) \in \mathbb{L}^2(\Omega)\right\}.$$

The domain of $\mathscr{A}$ is chosen to be a subset of its graph space:

$$\mathsf{D}(\mathscr{A}) := \{w \in \mathbb{H}_{\mathscr{A}} : u = 0 \text{ on } \partial\Omega\}$$

equipped with the graph norm $\|u\|_{\mathcal{G}} := \sqrt{\|u\|_{\mathbb{H}^1}^2 + \|\mathscr{A}u\|_{\mathbb{L}^2}^2} = \sqrt{\|u\|_{\mathbb{H}^1}^2 + \|\nabla \cdot (e^z \nabla u)\|_{\mathbb{L}^2}^2}$. Clearly, $\mathsf{D}(\mathscr{A})$ is dense in $\mathbb{L}^2(\Omega)$ and $\mathscr{A}$ is continuous on $\mathsf{D}(\mathscr{A})$. The proof of the self-adjointness of $\mathscr{A}$ is straightforward using basic facts from weak/distributional derivative and a standard distributional argument and it is provided in section 12.3 of section 12.3.

Example 12.8. We now consider a weak formulation for the elliptic differential operator in Example 12.7. This will be important for studying the well-posedness of the associated partial differential equations in Example 15.4. Multiplying $-\nabla \cdot (e^z \nabla u)$ by a test function and integrating by parts allow us to define the following bilinear form $a(\cdot,\cdot) : \mathbb{H}_0^1(\Omega) \times \mathbb{H}_0^1(\Omega) \to \mathbb{R}$

$$a(u,v) := (e^z \nabla u, \nabla v)_{\mathbb{L}^2(\Omega)}.$$

By the Cauchy-Schwarz inequality we have

$$|a(u,v)| \leq \|e^z\|_{\mathbb{L}^\infty(\Omega)} \|\nabla u\|_{\mathbb{L}^2(\Omega)} \|\nabla v\|_{\mathbb{L}^2(\Omega)} \leq \|e^z\|_{\mathbb{L}^\infty(\Omega)} \|u\|_{\mathbb{H}^1(\Omega)} \|v\|_{\mathbb{H}^1(\Omega)},$$

and thus $a(\cdot,\cdot)$ is continuous on $\mathbb{H}_0^1(\Omega) \times \mathbb{H}_0^1(\Omega)$. Here, $\mathbb{L}^\infty(\Omega)$ is the space of essentially bounded functions on Ω. As a result, it implicitly defines a unique linear and continuous operator[3] $\mathscr{A} : \mathbb{H}_0^1(\Omega) \to \mathbb{H}_0^1(\Omega)$ as

$$(\mathscr{A}u, v)_{\mathbb{H}_0^1(\Omega)} := a(u,v).$$

[3] Note that for every continuous sequilinear form $a : \mathbb{X} \times \mathbb{Y} \to \mathbb{F}$, there exists a unique linear and continuous operator $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ such that $(\mathscr{A}u, v)_{\mathbb{Y}} := a(u,v)$ for all $u \in \mathbb{X}$ and $v \in \mathbb{Y}$. Indeed, by the linearity and continuity of $a(\cdot,\cdot)$ with respect to its first argument, the Riesz representation Theorem 5.1 ensures that there exists a unique $\mathscr{A}u \in \mathbb{Y}$ such that $(\mathscr{A}u, v)_{\mathbb{Y}} = a(u,v)$. The continuity of $\mathscr{A}$ is from the continuity of $a(u,v)$: $\|\mathscr{A}u\| = \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A}u, v)}{\|v\|_{\mathbb{Y}}} = \sup_{v \in \mathbb{Y}} \frac{a(u,v)}{\|v\|_{\mathbb{Y}}} \leq \beta \|u\|_{\mathbb{X}}$. The uniqueness of $\mathscr{A}$ is straightforward due to its definition, as if there were another linear and continuous operator $\mathscr{B}$, then we would have

$$\|\mathscr{B} - \mathscr{A}\| = \sup_{u \in \mathbb{X}} \frac{\|\mathscr{B}u - \mathscr{A}u\|_{\mathbb{Y}}}{\|u\|_{\mathbb{X}}} = \sup_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{(\mathscr{B}u - \mathscr{A}u, v)_{\mathbb{Y}}}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} = \sup_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{a(u,v) - a(u,v)}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} = 0,$$

and thus $\mathscr{B} = \mathscr{A}$.

Due to the symmetry of the bilinear form, we have $\mathscr{A}^* = \mathscr{A}$.

Example 12.9. Consider an open and bounded domain $\Omega \subset \mathbb{R}^n$ and the $\mathbb{L}^2-$inner product of two functions $u(\boldsymbol{x}), v(\boldsymbol{x})$ is given in (12.5). Consider the following parametrized linear operator

$$\mathscr{A}u = \begin{cases} \boldsymbol{\beta} \cdot \nabla u + \lambda u & \text{in } \Omega, \\ \boldsymbol{\beta} \cdot \boldsymbol{n} u = 0 & \text{in } \partial\Omega_{\text{in}}, \end{cases}$$

where $\boldsymbol{\beta} \in \left[\mathcal{C}^1(\Omega, \mathbb{R})\right]^n$ and $\nabla \cdot \boldsymbol{\beta} = 0$, $\lambda > 0$, $\boldsymbol{n}$ is the unit outward normal vector of the boundary $\partial\Omega$, and $\partial\Omega_{\text{in}} := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{\beta} \cdot \boldsymbol{n} < 0\}$ is the inflow boundary. We consider the graph space $\mathbb{H}_{\mathscr{A}} := \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) := \left\{u : u \in \mathbb{L}^2(\Omega) \text{ and } \boldsymbol{\beta} \cdot \nabla u \in \mathbb{L}^2(\Omega)\right\}$ which is dense in $\mathbb{L}^2(\Omega)$ under sufficient regularity[4] of the domain Ω [52]. The domain of $\mathscr{A}$ is defined as a subset of the graph space, namely, $\mathsf{D}(\mathscr{A}) := \left\{u \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) : \boldsymbol{\beta} \cdot \boldsymbol{n} u = 0 \text{ in } \partial\Omega_{\text{in}}\right\}$. It is clear that $\mathscr{A} : \mathsf{D}(\mathscr{A}) \to \mathbb{L}^2(\Omega)$ is linear and continuous owing to the definition of $\mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)$ and the intrinsic graph norm $\|u\|_{\mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)} := \sqrt{\|u\|^2_{\mathbb{L}^2} + \|\boldsymbol{\beta} \cdot \nabla u\|^2}$. Using the definition of weak/distributional derivative and integration by parts once, we can show (see Example 12.10 for a more general differential operator) that

$$\mathscr{A}^* v = \begin{cases} -\boldsymbol{\beta} \cdot \nabla v + \lambda v & \text{in } \Omega, \\ \boldsymbol{\beta} \cdot \boldsymbol{n} v = 0 & \text{in } \partial\Omega_{\text{out}}, \end{cases}$$

where $\partial\Omega_{\text{out}} := \{\boldsymbol{x} \in \partial\Omega : \boldsymbol{\beta} \cdot \boldsymbol{n} > 0\}$ is the outflow boundary and

$$\mathsf{D}(\mathscr{A}^*) = \left\{v \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) : \boldsymbol{\beta} \cdot \boldsymbol{n} u = 0 \text{ in } \partial\Omega_{\text{out}}\right\}.$$

Example 12.10 (Friedrichs' systems). We are interested in Friedrichs' system that embraces a large class of elliptic, parabolic, hyperbolic, and mixed-type PDEs operators [58]:

$$\mathscr{A}u := \sum_{k=1}^{n} \boldsymbol{A}_k \partial_k u + \boldsymbol{C}u \text{ in } \ \in \Omega, \tag{12.6}$$

where d is the spatial dimension, u the unknown solution with values in $\mathbb{R}^m$, f the forcing term, and Ω is an open and bounded subset of $\mathbb{R}^n$ with sufficient regular boundary $\partial\Omega$. The matrices $\boldsymbol{A}_k$ and $\boldsymbol{C}$ are assumed to be constant and continuous across Ω. Here, ∂_k is understood as the kth partial derivative. We start with the standard assumptions (see, e.g., [58, 52, 78, 54]):

[4] Assume Ω has segment property [8].

$$\boldsymbol{C} \in \left[\mathbb{L}^{\infty}(\Omega)\right]^{m,m}, \tag{12.7a}$$

$$\boldsymbol{A}_k \in \left[\mathbb{L}^{\infty}(\Omega)\right]^{m,m}, \quad k=1,\ldots,n, \quad \text{and} \quad \sum_{k=1}^{n} \partial_k \boldsymbol{A}_k \in \left[L^{\infty}(\Omega)\right]^{m,m}, \tag{12.7b}$$

$$\boldsymbol{A}_k = (\boldsymbol{A}_k)^T \text{ in } \Omega, \quad k=1,\ldots,n,, \tag{12.7c}$$

$$\boldsymbol{C} + \boldsymbol{C}^T + \sum_{k=1}^{n} \partial_k \boldsymbol{A}_k \geq 2\alpha_0 I \text{ in } \Omega, \tag{12.7d}$$

where $\alpha_0 > 0$ is some coercivity constant. In this book, we consider the following abstract boundary condition for $\mathscr{A}$ [52]: let $\boldsymbol{D} := \sum_{k=1}^{n} \boldsymbol{A}_k \boldsymbol{n}_k$, where $\boldsymbol{n}_k$ is the kth component of the unit outward normal vector $\boldsymbol{n}$ on $\partial\Omega$, and assume there exists $\boldsymbol{M} \in \left[\mathbb{L}^{\infty}(\partial\Omega)\right]^{m,m}$ such that

$$\boldsymbol{M} + \boldsymbol{M}^T \geq 0 \text{ on } \partial\Omega, \tag{12.8a}$$

$$(\boldsymbol{D} - \boldsymbol{M})\, u = 0 \text{ on } \partial\Omega, \tag{12.8b}$$

$$\mathsf{N}(\boldsymbol{D} - \boldsymbol{M}) + \mathsf{N}(\boldsymbol{D} + \boldsymbol{M}) = \mathbb{R}^m, \text{ on } \partial\Omega. \tag{12.8c}$$

Following [51], we define[5] the graph space of $\mathscr{A}$ using its differential part $\mathscr{B} := \mathscr{A} - \boldsymbol{C}$:

$$\mathbb{H}_{\mathscr{A}} := \left\{ u \in \left[\mathbb{L}^2(\Omega)\right]^m : \mathscr{B}u \in \left[\mathbb{L}^2(\Omega)\right]^m \right\},$$

which is dense [54] in $\left[\mathbb{L}^2(\Omega)\right]^m$ when Ω is sufficiently regular (see Example 12.9 for an example). Furthermore, from the definition, it is easy to see that $\mathscr{A}$ is linear and continuous on $\mathsf{D}(\mathscr{A}) := \{x \in \mathbb{H}_{\mathscr{A}} : (\boldsymbol{D} - \boldsymbol{M})\, x = 0 \text{ on } \partial\Omega\}$ equipped with intrinsic graph norm. Using the definition of weak/distributional derivative and integration by parts we can show that its adjoint is found to be (see section 12.3 in section 12.3):

$$\mathscr{A}^* v = -\sum_{k=1}^{n} \partial_k (\boldsymbol{A}_k v) + \boldsymbol{C}^T v \in \mathbb{L}^2(\Omega),$$

for any $v \in \mathsf{D}(\mathscr{A}^*)$, where

$$\mathsf{D}(\mathscr{A}^*) = \left\{ \boldsymbol{v} \in \mathbb{H}_{\mathscr{A}} : \boldsymbol{D}\boldsymbol{v} \in [\mathsf{N}(\boldsymbol{D} - \boldsymbol{M})]^{\perp} \right\}.$$

[5] Note that we pick the $\mathbb{L}^2$-setting here for concreteness, but all the results for Friedrichs' system/operator in this paper hold for the general Hilbert space setting in [51, 52].

Remark 12.3. Note that we have considered Friedrichs systems with full coercivity in Example 12.10. Applying such a general setting to various concrete PDEs [51, 52] will reveal concrete adjoints and their domains, but we omit the details here. The elliptic operator in Example 12.7, when written in the first order form, is a particular two-field Friedrichs system [51, 52], whose adjoint can be derived similarly.

Problems

Problem 12.1 (Sturm-Liouville differential operator). Consider the following weighted $\mathbb{L}^2-$inner product of two functions u, v in $\mathbb{L}^2(a,b)$ over the complex field

$$(u,v)_{L^2(a,b)} = \int_a^b \rho(x)u(x)\overline{v(x)}dx, \tag{12.9}$$

where $\rho(x) \in \mathcal{C}[a,b]$ is a positive weight. Next, define an operator $\mathscr{A} : \mathsf{D}(\mathscr{A}) \subset \mathbb{L}^2(a,b) \to \mathbb{L}^2(a,b)$ as

$$\mathscr{A}u = \frac{1}{\rho}\left[-(pu')' + qu\right] \quad \text{in } (a,b)$$

where $p = p(x) \in \mathcal{C}^1[a,b]$ is positive, $q = q(x) \in \mathcal{C}[a,b]$, $()' = \dfrac{d}{dx}$. Let us choose the domain of $\mathscr{A}$ to be

$$\mathsf{D}(\mathscr{A}) := \left\{u \in \mathcal{C}^2[a,b] : \alpha_1 u(a) + \alpha_2 u'(a) = 0, \text{ and } \beta_1 u(b) + \beta_2 u'(b) = 0\right\},$$

where $\alpha_1, \alpha_2, \beta_1, \beta_2$ are constants such that $\alpha_1^2 + \alpha_2^2 \neq 0$ and $\beta_1^2 + \beta_2^2 \neq 0$. Find the adjoint and its domain using Definition 12.2.

Problem 12.2. Work out the details for Example 12.2.

Problem 12.3. Consider Example 12.1 but now the operator $\mathscr{A}$ is defined as

$$\mathscr{A}u := \frac{d^2u}{dx^2} + u.$$

We are interested in the domain of $\mathscr{A}$ with the boundary condition $u(1) = 0$. Define a suitable domain for $\mathscr{A}$, then find $\mathscr{A}^*$ (and its domain).

Problem 12.4. Let $\boldsymbol{v}$ be a function, with value in $\mathbb{R}^3$, over an open and bounded subset Ω of $\mathbb{R}^3$. Consider the Curl operator

$$\mathscr{A}\boldsymbol{v} := \nabla \times \boldsymbol{v}.$$

Pick $\mathsf{D}(\mathscr{A}) \subset \mathbb{L}^2(\Omega)$ such that $\mathscr{A}$ makes sense in the classical sense (i.e. ∇ is classical gradient operator). Then find $\mathscr{A}^*$ (and its domain).

12.3 Appendix

This section provides a proof for Example 12.7. It requires the notation of distributional derivatives which can be referred to Chapter 13.

Proof (of Example 12.7).
We next find $\mathscr{A}^*$ and $\mathsf{D}(\mathscr{A}^*)$. Take any $v \in \mathsf{D}(\mathscr{A}^*)$, by Definition 12.2 we have

$$\begin{aligned}
(\mathscr{A}u, v)_{\mathbb{L}^2(\Omega)} &= (u, \mathscr{A}^* v)_{\mathbb{L}^2(\Omega)}, \quad \forall u \in \mathsf{D}(\mathscr{A}) \\
&\Updownarrow \qquad\qquad\qquad\qquad \forall \varphi \in \mathcal{C}_0^\infty(\Omega) \text{ dense in } \mathsf{D}(\mathscr{A}) \\
(\mathscr{A}\varphi, v)_{\mathbb{L}^2(\Omega)} &= (\varphi, \mathscr{A}^* v)_{\mathbb{L}^2(\Omega)} \\
&\Updownarrow \qquad\qquad\qquad\qquad \text{definition of distributional derivative} \\
\langle \varphi, -\nabla \cdot (e^z \nabla v) \rangle &= (\varphi, \mathscr{A}^* v)_{\mathbb{L}^2(\Omega)} \\
&\Updownarrow \\
-\nabla \cdot (e^z \nabla v) &= \mathscr{A}^* v \in \mathbb{L}^2(\Omega)
\end{aligned}$$

which shows that $\mathscr{A}^* = \mathscr{A}$ on $\mathsf{D}(\mathscr{A}^*)$. Next, from

$$(\mathscr{A}u, v)_{\mathbb{L}^2(\Omega)} = (u, \mathscr{A}^* v)_{\mathbb{L}^2(\Omega)} = (u, \mathscr{A}v)_{\mathbb{L}^2(\Omega)}, \quad \forall u \in \mathsf{D}(\mathscr{A}) \text{ and } v \in \mathsf{D}(\mathscr{A}^*),$$

and integration by parts we conclude

$$-(e^z \nabla u \cdot \boldsymbol{n}, v)_{\mathbb{L}^2(\partial\Omega)} + (u, e^z \nabla v \cdot \boldsymbol{n})_{\mathbb{L}^2(\partial\Omega)} = 0, \quad \forall u \in \mathsf{D}(\mathscr{A}) \text{ and } v \in \mathsf{D}(\mathscr{A}^*),$$

with the assumption that the boundary integrals make sense. As a result, in addition to $\nabla \cdot (e^z \nabla v) \in \mathbb{L}^2$ we need $v \in \mathbb{H}^1(\Omega)$ and $v = 0$ on $\partial\Omega$. We conclude that $\mathsf{D}(\mathscr{A}^*) = \mathsf{D}(\mathscr{A})$, and $\mathscr{A}$ is self-adjoint.

Proof (Proof for Example 12.10). We now derive the adjoint $\mathscr{A}^*$ (and its domain) of the Friedrichs' operator in Example 12.10. Proceed as in section 12.3, we can easily see that

$$\mathscr{A}^* v = -\sum_{k=1}^{n} \partial_k (\boldsymbol{A}_k v) + \boldsymbol{C}^T v \in \mathbb{L}^2(\Omega),$$

for any $v \in \mathsf{D}(\mathscr{A}^*)$. Thus $\mathsf{D}(\mathscr{A}^*) \subseteq \mathbb{H}_{\mathscr{A}}$. Next from

$$(\mathscr{A}u, v)_{\mathbb{L}^2(\Omega)} = (u, \mathscr{A}^* v)_{\mathbb{L}^2(\Omega)}, \quad \forall u \in \mathsf{D}(\mathscr{A}) \text{ and } v \in \mathsf{D}(\mathscr{A}^*),$$

and integrating by parts, we obtain

$$(u, \boldsymbol{D}v)_{\mathbb{L}^2(\partial\Omega)} = 0, \quad \forall u \in \mathsf{D}(\mathscr{A}) \text{ and } v \in \mathsf{D}(\mathscr{A}^*),$$

which in turns yields $\boldsymbol{D}v \in \mathsf{N}(\boldsymbol{D} - \boldsymbol{M})^{\perp}$ since $u \in \mathsf{N}(\boldsymbol{D} - \boldsymbol{M})$. We thus conclude

$$\mathsf{D}(\mathscr{A}^*) = \left\{ \boldsymbol{v} \in \mathbb{H}_{\mathscr{A}} : \boldsymbol{D}\boldsymbol{v} \in [\mathsf{N}(\boldsymbol{D} - \boldsymbol{M})]^{\perp} \right\}$$

Chapter 13
Distributional derivatives and Green functions from adjoint perspectives

Abstract This chapter presents an introduction to distributional derivatives, Green functions, and Green identities. The goal is to provide fresh and constructive perspectives on these important subjects using adjoint. Using integration by parts, we shall show that adjoints of differential operators are also differential operators with appropriate boundary (and/or initial) conditions. Of particular interest are the derivations of the adjoint of the divergence, gradient, and curl operators. As a by-product, we shall see that by not specifying the boundary conditions for differential operators and their adjoints in the process of constructing the adjoints, we obtain the Green identities. The fact that the adjoint of a differential operator is a differential operator and that the adjoint of adjoint is the original operator has a far-reaching implication: we can generalize the notion of derivatives for functions that do not have classical derivatives. The two important ingredients that facilitate the generalization are: i) choose the domain of a differential operator $\mathscr{A}$ so that it makes sense in the classical sense; ii) use duality pairing to define the adjoint $\mathscr{A}^*$: $\langle \mathscr{A}u, v\rangle := \langle u, \mathscr{A}^* v\rangle$, where v is the function we would like to define its generalized derivatives. The beauty here is that if v possesses classical derivatives, we recover $\mathscr{A}^*$ as a differential operator in the classical sense. Otherwise, $\mathscr{A}^*$ is a generalized derivative known as distributional derivative or distribution. The special cases of distributions are weak derivatives that are extremely important for analyzing differential equations with Sobolev spaces. We shall see that Sobolev spaces are spaces of functions possessing up to certain order of weak derivatives. An important application of distribution is to rigorously understand the "Dirac delta function": in particular, it is not a function but a distribution. This is particularly useful as the Dirac delta distribution is the heart of the Green function theory. Unlike the standard literature where the Green function is defined through a differential equation with Dirac delta distributions as the right-hand side, we present a goal-oriented approach that systematically constructs Green functions equations via the Lagrangian theorem Theorem 9.3. We shall see that the Green function equation is nothing more but the adjoint equation. Our

approach, thus, provides not only a systematic way to the construction of Green functions for specific interest but also new insights into Green functions as the Lagrange multipliers. The prerequisites for this chapter are the following

- Basic familiarity with Lebesgue integration theory including equivalent classes of functions (functions that are the same almost everywhere) and the dominated convergence theorem.

13.1 Distribution and weak derivatives as adjoint of classical differential operators

We have seen in Chapter 12 that adjoints of classical differential operators are again differential operators. That was possible because we limited the domain of the adjoint operators to be subsets of spaces of continuously differentiable functions. Had we chosen "rougher" spaces of functions for the domain of the adjoint operator in section 12.2, we would not have been able, in the classical sense, to carry out the necessary integration by parts to obtain the adjoint explicitly. On the other hand, the domain of an abstract adjoint in Definition 12.2:

$$\mathsf{D}\left(\mathscr{A}^*\right) := \{v : \text{ the map } u \mapsto (\mathscr{A}u, v)_{\mathbb{V}} \text{ is continuous on } \mathbb{U}\},$$

says that as long as $(\mathscr{A}u, v)_{\mathbb{V}}$ is continuous in u with respect to the $\mathbb{U}$-norm, v is a member of $\mathsf{D}\left(\mathscr{A}^*\right)$ by definition. Moreover, section 12.1 shows that the adjoint $\mathscr{A}^*$ then exists, is unique, and is linear in v. In other words, we do not have to restrict ourselves in classical spaces of continuously differentiable functions to define the domain of adjoints of classical differentiable operators. But of course, the adjoint will change as we have discussed in section 12.1. Reversing this line of arguments provides us a way to generalize the classical derivatives to functions that are not differentiable in the classical sense. In particular, by enlarging $\mathsf{D}\left(\mathscr{A}^*\right)$ to include functions that are not classically differentiable, $\mathscr{A}^*$ is a generalized differentiable operator (interchangeably called a distributional derivative). What remains is to ensure that $(\mathscr{A}u, v)_{\mathbb{V}}$ be continuous with respect to u in the $\mathbb{U}$-topology. *The main goal of this section is to provide a brief introduction with basic ingredients to ensure that $\mathscr{A}^*$ is well-defined as a distributional derivative.* Before doing so, we need to answer a critical question: *why do we need to generalize the classical derivatives in the first place?*

13.1.1 Why do we need generalized derivatives?

In this section, we provide simple examples to present the need for a generalized notion of derivatives. *For the rest of the chapter, unless otherwise stated, when we write "differentiable" or "derivatives", we mean the classical derivatives.* Consider the simplest differential operator

$$\mathscr{A}u := u' := \frac{du}{dt}.$$

and we are interested in solving the most basic ordinary differentiable equation

$$\mathscr{A}u = f, \quad u(0) = u_0, \tag{13.1}$$

for $t \in (0, t_f)$, where the initial condition u_0 is given and f is left unspecified at this moment. Recall that $\mathcal{C}^n(\Omega)$ is the space of n-times continuously differentiable functions on a given domain Ω and the space of continuous functions $\mathcal{C}(\Omega)$ corresponds to $n = 0$. The obvious choice for the domain of $\mathscr{A}$ is $\mathsf{D}(\mathscr{A}) = \mathcal{C}^1(0, t_f)$, and thus $\mathsf{R}(\mathscr{A}) = \mathcal{C}(0, t_f)$: a proper subset of $\mathbb{L}^2(0, t_f)$ ($\mathbb{L}^2$ spaces and more generally $\mathbb{L}^p$ spaces will be defined formally in Definition 13.7). The main problem with this classical setting is that (13.1) does not have a solution if the forcing $f \in \mathbb{L}^2(0, t_f)$. For example, the forcing term f could be discontinuous, which is not uncommon in practice. The nature of this problem is that the domains of the definition of classical derivatives are too small to embrace a solution. Thus, there is a critical need to overcome this issue. The key idea is to define generalized differential operators with sufficient rich domains of definition that contain at least one solution. In particular, for (13.1) with $f \in \mathbb{L}^2(0, t_f)$, if the generalized differential operator $\mathscr{A}$ is defined such that its range is $\mathbb{L}^2(0, t_f)$, then we always have a solution.

In the following, subsection 13.1.2 introduces the space of distributions and we show that the Dirac delta is not a function but a distribution. Built upon this, we shall introduce the notion of distributional derivatives in subsection 13.1.4. Weak derivatives are then presented as special cases of distributational derivatives. As shall be seen, the Sobolev spaces are spaces of functions possessing certain order of weak derivatives.

13.1.2 The space of distributions

As discussed in Remark 12.2, enlarging $\mathsf{D}(\mathscr{A}^*)$ means restricting $\mathsf{D}(\mathscr{A})$ to a smaller subspace. One of the sufficiently small spaces is the space of continuously infinitely differentiable functions with compact supports.

Definition 13.1 (compact subsets of $\mathbb{R}^n$). A set K is a compact subset of $\mathbb{R}^n$ if it is closed and bounded, that is,

- $\overline{\mathsf{K}} = \mathsf{K}$, and
- there exists $0 < \alpha < \infty$ such that $\|\boldsymbol{x}\|_{\mathbb{R}^n} < \alpha$ for all $\boldsymbol{x} \in \mathsf{K}$.

Definition 13.2 (Support of a function). Let $\Omega \subseteq \mathbb{R}^n$ and $f : \Omega \to \mathbb{F}$ (either real or complex field). We define a support of f on Ω as

$$\operatorname{supp} f = \overline{\{x \in \Omega : f(x) \neq 0\}},$$

where the overline denotes the closer (recall Definition 5.8) in $\mathbb{R}^n$.

Definition 13.3 (Continuously infinitely differentiable function). We define

$$\mathcal{D}(\Omega) := \mathcal{C}_0^\infty(\Omega) := \{f : f \text{ is continuously infinitely differentiable functions and } \operatorname{supp} f \text{ is a compact subset of } \Omega\} \quad (13.2)$$

Note that we have introduced two alternative notations $\mathcal{D}(\Omega)$ and $\mathcal{C}_0^\infty(\Omega)$ for the space of continuously infinitely differentiable functions with compact support in Ω. *We shall interchangeably call $\mathcal{D}(\Omega)$ as the test space (on Ω), and each member of which is called a test function.*

We next introduce the multi-index notation that helps us simplify the notations for mixed derivatives in high dimensions.

Definition 13.4 (Multi-index notation for mixed derivatives). A multi-index $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \dots \boldsymbol{\alpha}_n]^T$ is an n-tuple of non-negative integers. The norm of multi-index is defined as:

$$|\boldsymbol{\alpha}| := \sum_{i=1}^{n} \boldsymbol{\alpha}_i.$$

Let $\mathscr{D}$ be a derivative notation: it could the classical, or Fréchet (see Chapter 9), or the distributional derivative (to be defined). We define te mixed derivative of order $|\boldsymbol{\alpha}|$ for a function[1] $u : \Omega \subseteq \mathbb{R}^n \to \mathbb{F}$, where $\mathbb{F}$ is either the complex or real field, as

$$\mathscr{D}^{\boldsymbol{\alpha}} u = \mathscr{D}_{\boldsymbol{x}_1}^{\boldsymbol{\alpha}_1} \mathscr{D}_{\boldsymbol{x}_2}^{\boldsymbol{\alpha}_2} \dots \mathscr{D}_{\boldsymbol{x}_n}^{\boldsymbol{\alpha}_n} u.$$

If $|\boldsymbol{\alpha}| = 0$, we set

$$\mathscr{D}^{\boldsymbol{\alpha}} u := u.$$

Example 13.1. Let $u : \mathbb{R}^3 \to \mathbb{R}$ and $\boldsymbol{\alpha} = [\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \boldsymbol{\alpha}_3]^T$ be an multi-index. Assume that u is differentiable up to order $|\boldsymbol{\alpha}|$. Using ∂ for classical partial derivatives, we have

$$\partial^{\boldsymbol{\alpha}} u = \frac{\partial^{\boldsymbol{\alpha}_1}}{\partial \boldsymbol{x}_1^{\boldsymbol{\alpha}_1}} \left(\frac{\partial^{\boldsymbol{\alpha}_2}}{\partial \boldsymbol{x}_2^{\boldsymbol{\alpha}_2}} \left(\frac{\partial^{\boldsymbol{\alpha}_3}}{\partial \boldsymbol{x}_3^{\boldsymbol{\alpha}_3}} u \right) \right).$$

[1] generally a distribution.

For example, taking $\alpha = [1,\ 2,\ 3]^T$ gives

$$\partial^{\boldsymbol{\alpha}} u = \frac{\partial}{\partial \boldsymbol{x}_1}\left(\frac{\partial^2}{\partial \boldsymbol{x}_2^2}\left(\frac{\partial^3}{\partial \boldsymbol{x}_3^3} u\right)\right) = \frac{\partial^6}{\partial \boldsymbol{x}_1 \partial \boldsymbol{x}_2^2 \partial \boldsymbol{x}_3^3} u.$$

Definition 13.5 (Uniform norm). Let $\Omega \subset \mathbb{R}^n$ be open and bounded, we define the (uniform) norm in the space of continuous functions on Ω, $\mathcal{C}(\Omega)$, as

$$\|f\|_\infty := \|f\|_{\mathcal{C}(\Omega)} := \sup_{\boldsymbol{x}\in\Omega} |f(\boldsymbol{x})|, \quad \forall f \in \mathcal{C}(\Omega).$$

Definition 13.6 (Convergence in $\mathcal{D}(\Omega)$). A sequence $\{\phi_n\}_{n=1}^\infty \subset \mathcal{D}(\Omega)$ converges to $\phi \in \mathcal{D}(\Omega)$ iff

i) there exists a compact subset $\mathsf{K} \subset \Omega$ such that $\operatorname{supp}\phi_n \subset \mathsf{K}, \forall n$, and
ii) $\|\partial^{\boldsymbol{\alpha}}\phi_n - \partial^{\boldsymbol{\alpha}}\phi\|_{\mathcal{C}(\mathsf{K})} \to 0$ for all multi-index $\boldsymbol{\alpha}$.

Remark 13.1. Note that any derivative of a test function is again a test function. We will use this fact at various places in subsection 13.1.4.

Remark 13.2. Though all the results, definitions, etc, when appropriate, can be extended straightforwardly to complex-value functions and/or vector spaces over complex fields, we shall only consider the real cases for clarity.

For the rest of the chapter, $\int d\boldsymbol{x}$ means a Lebesgue integral with $d\boldsymbol{x}$ as the Lebesgue measure. Unless otherwise stated, we assume $\Omega \subseteq \mathbb{R}^n$. We first define $\mathbb{L}^p(\Omega)$ spaces here for any natural number $p = 1, 2, \ldots$, as we will refer to them frequently from now on.

Definition 13.7 ($\mathbb{L}^p$ spaces). Let p be an arbitrary natural number and $\Omega \subseteq \mathbb{R}^d$, we define the corresponding $\mathbb{L}^p(\Omega)$ space as

$$\mathbb{L}^p(\Omega) := \left\{ f : \mathbb{R}^d \to \mathbb{F} \text{ such that } \|f\|_{\mathbb{L}^p(\Omega)} := \left(\int_\Omega |f(\boldsymbol{x})|^p \, d\boldsymbol{x}\right)^{\frac{1}{p}} < \infty \right\},$$

where $\mathbb{F}$ is either real or complex field and $\|f\|_{\mathbb{L}^p(\Omega)}$ is called the $\mathbb{L}^p$-norm of f. The generalization to vector-valued $\mathbb{L}^p$-spaces is straightforward:

$$[\mathbb{L}^p(\Omega)]^n := \left\{ \boldsymbol{f} : \mathbb{R}^d \to \mathbb{F}^n \text{ such that } \|\boldsymbol{f}\|_{[\mathbb{L}^p(\Omega)]^n} := \left(\int_\Omega \|\boldsymbol{f}(\boldsymbol{x})\|_{\ell^p}^p \, d\boldsymbol{x}\right)^{\frac{1}{p}} < \infty \right\},$$

where the ℓ^p-norm in $\mathbb{R}^n$ is defined as

$$\|\boldsymbol{f}(\boldsymbol{x})\|_{\ell^p} := \left(\sum_{i=1}^p |\boldsymbol{f}_i(\boldsymbol{x})|^p\right)^{\frac{1}{p}}.$$

When $p = \infty$ we define $\mathbb{L}^{\infty}(\Omega)$ slightly differently.

$$\mathbb{L}^{\infty}(\Omega) := \left\{ f : \mathbb{R}^d \to \mathbb{R} \text{ such that } \|f\|_{\mathbb{L}^{\infty}(\Omega)} := \operatorname{ess\,sup} |f| < \infty \right\},$$

where[2]

$$\operatorname{ess\,sup} f := \inf \{ \sup g : g = f \text{ almost everywhere} \}.$$

We say that $\mathbb{L}^{\infty}(\Omega)$ is the space of essentially bounded functions. The generalization to vector-valued function is obvious

$$[\mathbb{L}^{\infty}(\Omega)]^n := \left\{ \boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}^n \text{ such that } \|\boldsymbol{f}\|_{[\mathbb{L}^{\infty}(\Omega)]^n} := \max_{i=1\ldots,n} \operatorname{ess\,sup} \boldsymbol{f}_i < \infty \right\}.$$

These $\mathbb{L}^p$-spaces are Banach spaces and many of their properties can be referred to, e.g., [3, 102, 112]. For example

$$\mathcal{D}(\Omega) \underset{\text{dense}}{\subset} \mathbb{L}^p(\Omega), \quad \forall p \geq 1, \tag{13.3}$$

where $\underset{\text{dense}}{\subset}$ means "is a dense subset set of".

Remark 13.3 (A fact). $\mathbb{L}^2(\Omega)$ is a Hilbert space with the following inner product

$$(f, g)_{\mathbb{L}^2(\Omega)} := \int_{\Omega} f(\boldsymbol{x}) \overline{g(\boldsymbol{x})}\, d\boldsymbol{x},$$

for any $f, g \in \mathbb{L}^2(\Omega)$. For real-valued $\mathbb{L}^2$ spaces, the above $\mathbb{L}^2$-inner product becomes

$$(f, g)_{\mathbb{L}^2(\Omega)} := \int_{\Omega} f(\boldsymbol{x}) g(\boldsymbol{x})\, d\boldsymbol{x}.$$

We shall deploy the Cauchy-Schwarz inequality for $\mathbb{L}^2(\Omega)$ at various places in the book:

$$\left| (f, g)_{\mathbb{L}^2(\Omega)} \right| \leq \|f\|_{\mathbb{L}^2(\Omega)} \|g\|_{\mathbb{L}^2(\Omega)}, \quad \forall f, g \in \mathbb{L}^2(\Omega). \tag{13.4}$$

The next result generalizes those in section 12.2. In particular, it essentially shows that $\partial^{\boldsymbol{\alpha}}$ is, up to a sign, the adjoint of $\partial^{\boldsymbol{\alpha}}$. More importantly, it is the motivation to define distributional derivatives in Definition 13.12.

Proposition 13.1 ($(-1)^{|\boldsymbol{\alpha}|}\partial^{\boldsymbol{\alpha}}$ is the adjoint of $\partial^{\boldsymbol{\alpha}}$). *Consider $\Omega \subseteq \mathbb{R}^n$, $\phi \in \mathcal{D}(\Omega)$, a multi-index $\boldsymbol{\alpha}$, and $u : \mathbb{R}^n \to \mathbb{R}$ is $|\boldsymbol{\alpha}|$-order differentiable. Then*

$$\left((-1)^{|\boldsymbol{\alpha}|} \partial^{\boldsymbol{\alpha}} u, \phi \right)_{\mathbb{L}^2(\Omega)} = (u, \partial^{\boldsymbol{\alpha}} \phi)_{\mathbb{L}^2(\Omega)}. \tag{13.5}$$

[2] $g = f$ almost everywhere means f is equal g pointwise except on the sets with zero (Lebesgue) measure.

Proof. The proof is straightforward using integration by parts (see Problem 13.2).

Definition 13.8 (The space of distributions). Let $\Omega \subseteq \mathbb{R}^n$. The dual of $\mathcal{D}(\Omega)$, denoted as $\mathcal{D}'(\Omega)$, is the collection of linear functionals, called distributions (also known as generalized functions), u such that for every compact subset K of Ω, there exists a non-negative integer j and a positive constant c such that:

$$u(\phi) := \langle u, \phi \rangle \leq c\, P_{\mathsf{K},j}(\phi), \quad \text{for any } \phi \in \mathcal{D}(\Omega) \text{ with } \operatorname{supp}\phi \subset \mathsf{K},$$

where we have used $\langle u, \phi \rangle$ and $u(\phi)$ interchangeably for the duality pairings between $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$, and we have defined

$$P_{\mathsf{K},j}(\phi) := \sup_{|\boldsymbol{\alpha}| \leq j,\ \boldsymbol{x} \in \mathsf{K}} |\partial^{\boldsymbol{\alpha}} \phi(\boldsymbol{x})|, \tag{13.6}$$

with $\partial^{\boldsymbol{\alpha}}\phi$ denoting the $|\boldsymbol{\alpha}|$-order classical mixed derivative of ϕ.

It turns out that any $u \in \mathcal{D}'(\Omega)$ is not only linear but also (sequentially) continuous on $\mathcal{D}(\Omega)$.

Corollary 13.1. *$\mathcal{D}'(\Omega)$ is the collection of all (sequentially) continuous linear functionals on $\mathcal{D}(\Omega)$ in the following sense: if $\{\phi_n\}_{n=1}^{\infty} \subset \mathcal{D}(\Omega)$ converges to ϕ in $\mathcal{D}(\Omega)$, then for any $u \in \mathcal{D}'(\Omega)$ we have $u(\phi_n) \to u(\phi)$.*

Proof. Let $\{\phi_n\}_{n=1}^{\infty} \subset \mathcal{D}(\Omega)$ converges to ϕ in $\mathcal{D}(\Omega)$. Clearly $\phi_n - \phi \in \mathcal{D}(\Omega)$ and by Definition 13.6, there exists a compact subset K such that $\operatorname{supp}\phi \subset \mathsf{K}$ and $\operatorname{supp}\phi_n \subset \mathsf{K}$ for all n. Now, by Definition 13.8, for any $u \in \mathcal{D}'(\Omega)$, there exists a non-negative integer $j(n)$ and a constant $c(n)$ such that

$$u(\phi_n - \phi) \leq c(n) \sup_{|\boldsymbol{\alpha}| \leq j(n),\ \boldsymbol{x} \in \mathsf{K}} |\partial^{\boldsymbol{\alpha}}(\phi_n(\boldsymbol{x}) - \phi(\boldsymbol{x}))| \xrightarrow[n\to\infty]{\text{Definition 13.6}} 0,$$

and this concludes the proof.

Remark 13.4. It is important to note that both j and c in Definition 13.12 depend on the compact set K under consideration.

Example 13.2. In (13.6), if we take $j = 1$, then there are $n + 1$ possibilities for $\partial^{\boldsymbol{\alpha}}\phi$

$$\partial^{\boldsymbol{\alpha}}\phi = \left\{ \frac{\partial \phi}{\partial \boldsymbol{x}_i} : i = 1, \ldots, n \right\} \cup \{\phi\}.$$

Definition 13.9 (Locally integrable functions). If $\int_{\mathsf{K}} |f(\boldsymbol{x})|\, d\boldsymbol{x} < \infty$ for any compact subset K of $\Omega \subseteq \mathbb{R}^n$, then f is called locally integrable, and we denote by $\mathbb{L}^1_{loc}(\Omega)$ the collections of all locally integrable functions.

We can define similarly the space of locally square integrable functions $\mathbb{L}^p_{loc}(\Omega)$ as the spaces of functions whose $\mathbb{L}_p$-norm on any compact subset of Ω is finite. The following are two examples of subsets of $\mathbb{L}^1_{loc}(\Omega)$.

Corollary 13.2. *$\mathbb{L}^2_{loc}(\Omega) \subset \mathbb{L}^1_{loc}(\Omega)$ and $\mathcal{C}(\Omega) \subset \mathbb{L}^1_{loc}(\Omega)$. In particular, we have $\mathcal{C}^{\boldsymbol{\alpha}}(\Omega) \subset \mathbb{L}^1_{loc}(\Omega)$ for any multi-index $\boldsymbol{\alpha}$.*

Proof. For the first assertion, take any $f \in \mathbb{L}^2(\Omega)$, we have

$$\int_{\mathsf{K}} |f(\boldsymbol{x})| \, d\boldsymbol{x} \leq \sqrt{|\mathsf{K}|} \, \|f\|_{\mathbb{L}^2(\mathsf{K})} < \infty,$$

where we have used Cauchy-Schwarz inequality (13.4) in the first inequality and by $|\mathsf{K}|$ we mean the (Lebesgue) measure of the set K. For the second assertion, any continuous function f on Ω is locally integrable since

$$\|f\|_{\mathbb{L}^2(\mathsf{K})} \leq \max_{\boldsymbol{x} \in \mathsf{K}} |f(\boldsymbol{x})| \sqrt{|\mathsf{K}|} < \infty,$$

where $\max_{\boldsymbol{x} \in \mathsf{K}} |f(\boldsymbol{x})|$ is well-defined and finite owing to the Weierstrass theorem (i.e. any continuous function on a compact set attains its maximum [112, 109, 28, 12, 87, 142]).

Definition 13.10 (Regular distribution). For any $f \in \mathbb{L}^1_{loc}(\Omega)$, where $\Omega \subseteq \mathbb{R}^n$, we define the associate linear functional $\mathcal{F} : \mathcal{D}(\Omega) \mapsto \mathbb{R}$ (with an obvious extension to $\mathbb{C}$) as:

$$< \mathcal{F}, \phi > := \int_{\Omega} f \; \phi \; d\boldsymbol{x} = (f, \phi)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega), \tag{13.7}$$

then $\mathcal{F}$ is call a regular distribution generated/induced via the $\mathbb{L}^2$-inner product by the locally integrable function f.

We need to provide the proof for the claim that $\mathcal{F}$ in (13.7) is a distribution. Indeed, for any compact subset K of Ω and any $\phi \in \mathcal{D}(\Omega)$ with $\operatorname{supp} \phi \subset \mathsf{K}$ we have

$$\begin{aligned} |< \mathcal{F}, \phi >| = \left| \int_{\Omega} f(\boldsymbol{x}) \;\; \phi(\boldsymbol{x}) \; d\boldsymbol{x} \right| = \left| \int_{\mathsf{K}} f(\boldsymbol{x}) \;\; \phi(\boldsymbol{x}) \; d\boldsymbol{x} \right| \leq \int_{\mathsf{K}} |f(\boldsymbol{x})| \, |\phi(\boldsymbol{x})| \; d\boldsymbol{x} \\ \leq \max_{\boldsymbol{x} \in \mathsf{K}} |\phi(\boldsymbol{x})| \int_{\mathsf{K}} |f(\boldsymbol{x})| \; d\boldsymbol{x} = c \, P_{\mathsf{K},j}(\phi), \end{aligned}$$

with $j = 0$ and $c = \int_{\mathsf{K}} |f(\boldsymbol{x})| \; d\boldsymbol{x}$. Therefore, $\mathcal{F} \in \mathcal{D}'(\Omega)$ is a distribution by Definition 13.8

Remark 13.5. For $\mathcal{F} \in \mathcal{D}'(\Omega)$ that is generated from a locally integrable function $f \in \mathbb{L}^1_{loc}(\Omega)$ function, we simply say that $\mathcal{F}$ is a regular distribution.

It is conventional and convenient to identify $\mathcal{F}$ *with* f, and thus we simply write

$$< f, \phi >:= \int_{\Omega} f(\boldsymbol{x})\ \phi(\boldsymbol{x})\ d\boldsymbol{x} = (f, \phi)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega),$$

and call f itself as a regular distribution in this case. Such an identification makes sense as far as the test space is concerned since the action of $\mathcal{F}$ on any test function is the same as the $\mathbb{L}^2$-inner product of f with ϕ by definition. This is similar to the identification of the dual $\mathbb{X}^*$ of a Hilbert space $\mathbb{X}$ via the Riesz representation Theorem 5.1 (see Remark 5.4). In the following, we will identify a distribution with its generator via other means such as $\mathbb{H}^m$-inner product in the proof of Proposition 13.3. When it is important to re-emphasize this point for clarity, we will do so.

Remark 13.6. It is important to emphasize that a regular distribution $\mathcal{F}$ *is a special distribution whose action on a test function (i.e. pairing with a test function) can be written as the* $\mathbb{L}^2$*-inner product of a locally integrable function* f *and the test function.*

Lemma 13.1 (Uniqueness of regular distribution). *Let* $h \in \mathbb{L}^1_{loc}(\Omega)$ *and suppose:*

$$(h, \phi)_{\mathbb{L}^2(\Omega)} = 0, \quad \forall \phi \in \mathcal{D}(\Omega),$$

then $h = 0$ *almost everywhere in* Ω*. Consequently, if there are two regular distributions* $\mathcal{F}$ *and* G *such that their actions on the test space* $\mathcal{D}(\Omega)$ *are identical, then* $\mathcal{F} = G$*.*

Proof. The first assertion is the standard result for $\mathbb{L}^1_{loc}(\Omega)$ (see, e.g., [27]). For the second assertion, suppose $\mathcal{F}, G$ are induced by $f, g \in \mathbb{L}^1_{loc}(\Omega)$. We have

$$\langle \mathcal{F} - G, \phi \rangle = (f - g, \phi)_{\mathbb{L}^2(\Omega)} = 0, \quad \forall \phi \in \mathcal{D}(\Omega),$$

which means that $f = g$ almost everywhere by the first assertion, and thus $\mathcal{F} = G$.

Example 13.3. Consider the Heaviside function:

$$f(x) = \begin{cases} 0, & x \leq 0, \\ 1, & x > 0. \end{cases}$$

Clearly $f \in \mathbb{L}^1_{loc}(\mathbb{R})$. Now by (13.7), the regular distribution $\mathcal{F}$ generated by f is given by

$$< \mathcal{F}, \phi >= \int_{\mathbb{R}} f(x)\ \phi(x)\ dx = \int_0^{\infty} \phi(x)\ dx, \quad \forall \phi \in \mathcal{D}(\mathbb{R}).$$

13.1.3 Dirac delta is a distribution

We start by pointing out that our elementary understanding of the Dirac delta $\delta(\boldsymbol{x})$

$$\delta(\boldsymbol{x}) = \begin{cases} 0, & \boldsymbol{x} \neq \mathbf{0}, \\ \infty, & \boldsymbol{x} = \mathbf{0}. \end{cases}$$

is not useful under the Lebesgue measure because it is equivalent to the zero function. In other words, $\delta(\boldsymbol{x})$ is the zero function (in fact the equivalent class of zero function) in any $\mathbb{L}^p$-spaces in Definition 13.7.

Another notation of the Dirac delta that we also have seen is

$$\int_{\mathbb{R}} \delta(\boldsymbol{x})\ \phi(\boldsymbol{x})\ d\boldsymbol{x} = \phi(\mathbf{0}). \tag{13.8}$$

where $\phi \in \mathcal{D}(\Omega)$.

We first show that the definition in (13.8) does not make sense in $\mathbb{L}^1(\mathbb{R}^n)$. For simplicity, we only provide the proof for $\mathbb{L}^1(\mathbb{R})$. Suppose $\delta(x)$ resides in $\mathbb{L}^1(\mathbb{R})$, i.e $\int_{\mathbb{R}} |\delta(x)|\ dx < \infty$. If we take $\phi(x) \in \mathcal{D}(\Omega)$, where $\Omega = (-a, a) \subset \mathbb{R}$ for some $a > 0$, as:

$$\phi(x) = \begin{cases} e^{\frac{a^2}{x^2 - a^2}}, & |x| \leq a, \\ 0, & |x| \geq a, \end{cases} \tag{13.9}$$

we have

$$\begin{aligned} |\phi(0)| = \left| \int_{\mathbb{R}} \delta(x)\ \phi(x)\ dx \right| = \left| \int_{-a}^{a} \delta(x)\ \phi(x)\ dx \right| &\leq \int_{-a}^{a} |\delta(x)|\, |\phi(x)|\ dx \\ &\leq \sup_{|y| \leq a} |\phi(y)| \int_{-a}^{a} |\delta(x)|\ dx \leq e^{-1} \int_{-a}^{a} |\delta(x)|\ dx, \end{aligned} \tag{13.10}$$

where every step makes sense as $\sup_{|y| \leq a} |\phi(y)| = \phi(0) = e^{-1}$, and by the assumption

$$\int_{-a}^{a} |\delta(x)|\ dx \leq \int_{\mathbb{R}} |\delta(x)|\ dx < \infty.$$

Thus, (13.10) is equivalent to

$$e^{-1} \leq e^{-1} \int_{-a}^{a} |\delta(x)|\ dx \xrightarrow{a \to \infty} e^{-1} \times 0 = 0$$

due to the property of the Lebesque integral, which is a contradiction. We conclude that $\delta(x)$ is not a member of $\mathbb{L}^1(\mathbb{R})$, that is, it is not integrable.

We next show that the definition in (13.8) does not make sense in $\mathbb{L}^2(\mathbb{R}^n)$. For simplicity, we only provide the proof for $\mathbb{L}^2(\mathbb{R})$. Suppose $\delta(x)$ resides in

$\mathbb{L}^2(\mathbb{R})$, i.e $\int_{\mathbb{R}} |\delta(x)|^2 \, dx < \infty$ and take $\Omega = (-a, a)$. For any $\phi \in \mathcal{D}(\Omega)$, we have

$$|\phi(0)| = \left| \int_{\mathbb{R}} \delta(x) \; \phi(x) \; dx \right| = \left| \int_{-a}^{a} \delta(x) \; \phi(x) \; dx \right| \leq \sqrt{\int_{-a}^{a} |\delta(x)|^2 \; dx} \sqrt{\int_{-a}^{a} |\phi(x)|^2 \; dx},$$

by Cauchy Schwarz inequality (13.4). Now taking ϕ as in (13.9) leads to

$$|\phi(0)| \leq \sqrt{\int_{-a}^{a} |\delta(x)|^2 \; dx} \sqrt{\int_{-a}^{a} \left| \max_{|y| \leq a} \phi(y) \right|^2 \; dx} \leq e^{-1} \sqrt{2a} \sqrt{\int_{-a}^{a} |\delta(x)|^2 \; dx}.$$

Now passing the limit $a \to 0$ we have a contradiction:

$$e^{-1} \leq e^{-1} \times 0 \times 0 = 0.$$

Therefore, $\delta(x) \notin \mathbb{L}^2(\mathbb{R})$, i.e., it cannot be square integrable either.

In fact, we can show that the Dirac delta defined as in (13.8) cannot be a member of any $\mathbb{L}^p$ function spaces (see Problem 13.1). So which space does it belong to? To answer this question, we need a proper definition for $\delta(\boldsymbol{x})$.

Definition 13.11 (Dirac delta). Let $\Omega \subseteq \mathbb{R}^n$ such that $\mathbf{0} \in \Omega$. We define δ via the following action on the test space

$$\langle \delta, \phi \rangle := \phi(\mathbf{0}), \quad \forall \phi \in \mathcal{D}(\Omega). \tag{13.11}$$

Proposition 13.2 (Dirac delta is a distribution). *The Dirac delta δ defined in Definition 13.11 is a distribution, i.e., $\delta \in \mathcal{D}'(\mathbb{R}^n)$.*

Proof. For any compact subset K of $\mathbb{R}^n$ we have

$$|\langle \delta, \phi \rangle| = |\phi(\mathbf{0})| \leq \max_{\boldsymbol{x} \in \mathsf{K}} |\phi(\boldsymbol{x})| = P_{\mathsf{K},0}(\phi), \quad \forall \phi \in D(\mathbb{R}^n) \text{ with } \operatorname{supp} \phi \subset \mathsf{K}.$$

Hence by Definition 13.8, $\delta \in \mathcal{D}'(\mathbb{R}^n)$, i.e., δ is a distribution.

In subsection 13.1.4, we shall see a clearer picture for δ at least in $\mathbb{R}$, in which we show that δ is the distributional derivative of the Heaviside function. In Corollary 13.6, we shall precisely show that δ is a linear continuous functional on a Sobolev space and provide an explicit expression of δ via its Riesz representation.

13.1.4 Distributional derivatives as generalized derivatives

As discussed in Remark 12.2 and at the beginning of section 13.1, in order to use the adjoint differential operator $\mathscr{A}^*$ to define generalized derivatives for rougher functions, we need to enlarge $\mathsf{D}(\mathscr{A}^*)$ and simultaneously restrict the domain $\mathsf{D}(\mathscr{A})$ of the classical differential operator $\mathscr{A}$. The theory of distributions, due to Schwartz [129, 102, 11, 141], take the smoothnest space of functions $\mathcal{D}(\Omega)$ as $\mathsf{D}(\mathscr{A})$ and take the roughest space of "functions" (distributions to be precise) $\mathcal{D}'(\Omega)$ as $\mathsf{D}(\mathscr{A}^*)$. The adjoint differential operator $\mathscr{A}^*$ is then called distributional derivatives. It is a generalized notation of the classical derivative in the sense that when $\mathscr{A}^*$ acts on functions that possesses a classical derivative of the same order, $\mathscr{A}^*$ reduces to that classical derivative.

Before dwelling into the definition of distributional derivatives as $\mathscr{A}^*$, *we need to resolve one technical issue*: we have defined $\mathscr{A}^*$ as the adjoint of an operator $\mathscr{A}$ mapping between two Hilbert spaces, but it is not clear if $\mathsf{D}(\mathscr{A}^*) := \mathcal{D}'(\Omega)$ is a (dense) subset of any Hilbert space, and thus defining $\mathscr{A}^*$ via inner products may not be feasible. The theory of distribution address this by defining the adjoint of classical differential operators using duality pairing $\langle \cdot, \cdot \rangle$ between $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$.

Definition 13.12 (Adjoint of classical differential operators via duality pairings). Let $\boldsymbol{\alpha}$ be a multi-index and u be a distribution, we call $\mathscr{D}^{\boldsymbol{\alpha}}$ with the domain $\mathcal{D}'(\Omega)$ the "adjoint" of $\partial^{\boldsymbol{\alpha}}$ with the domain $\mathcal{D}(\Omega)$ if

$$\left\langle (-1)^{|\boldsymbol{\alpha}|} \mathscr{D}^{\boldsymbol{\alpha}} u,\ \phi \right\rangle = \langle u,\ \partial^{\boldsymbol{\alpha}} \phi \rangle\,, \quad \forall \phi \in \mathcal{D}(\Omega) \text{ and } u \in \mathcal{D}'(\Omega)\,. \tag{13.12}$$

$\mathscr{D}^{\boldsymbol{\alpha}}$ is called the distributional derivatives of u.

A few remarks for Definition 13.12 are in order. First, we note the resemblance between (13.12) and the integration by parts formula (13.5) for differentiable functions. In fact the latter is the motivation for the former. This ensures the generality and consistency of the distributional derivatives in the sense that when u is indeed classically differentiable, its distributional derivatives coincide with its classical derivatives (see Example 13.5). Second, we use the duality pairing $\langle \cdot, \cdot \rangle$ between $\mathcal{D}(\Omega)$ and $\mathcal{D}'(\Omega)$ to define the "adjoint derivatives" $\mathscr{D}^{\boldsymbol{\alpha}} u$.[3] If we ignore the factor $(-1)^{|\boldsymbol{\alpha}|}$ (either 1 or -1, and

[3] Using duality pairings to define adjoint is standard for a mapping between two Banach spaces $\mathbb{U}$ and $\mathbb{V}$.

$$\langle \mathscr{A} u, v \rangle_{\mathbb{V}^* \times \mathbb{V}} = \langle \mathscr{A}^* v, u \rangle_{\mathbb{U}^* \times \mathbb{U}}\,,$$

where $\mathbb{U}^*$ and $\mathbb{V}^*$ are the duals of $\mathbb{U}$ and $\mathbb{V}$, respectively. When $\mathbb{U}$ and $\mathbb{V}$ are Hilbert spaces, this reduces to the standard adjoint definition Definition 12.2 as we can identify a Hilbert space and its dual (see the discussion on this identification and the Riesz map right after Definition 5.5). The rigorous identification of the adjoint is via the Riesz map, but we ignore the details here.

thus amounting to a sign change), the above adjoint operator $\mathscr{D}^{\boldsymbol{\alpha}}$ is known as the transpose operator in some literature due to the duality pairings instead of inner products[4] (see [112] for example). Third, this definition of the adjoint $\mathscr{D}^{\boldsymbol{\alpha}}$ is different up to a negative sign compared to the standard definition Definition 12.2. In the literature, (13.12) is used to defined distributional derivatives without making connection to adjoint (or transpose operator). *Our goal in this section is to use adjoint to provide a personal insight into the origin of distributional derivatives and their definitions despite of the danger of not being entirely precise.* The gain is however worthwhile: our constructive derivation of distributional derivatives from adjoint perspective helps the readers (provided that they have some familiarity with adjoint) connects the gap between the distributional derivatives and adjoint, and thus providing a continuation of learning with insights and depth.

Fourth, Definition 13.12 makes sense. In particular, a direct consequence is that the distributional derivative of any order is again a well-defined distribution. This is in contrast to the classical derivative, the existence of which depends on the regularity of the function under consideration.

Corollary 13.3 (Distributional derivatives of distributions are again distributions). *The distributional derivative $\mathscr{D}^{\boldsymbol{\alpha}}u$ defined in (13.12) is a distribution.*

Proof. Since u is a distribution, from Definition 13.8, for every compact subset K of Ω, there exists a non-negative integer j and a positive constant c such that:

$$|\langle u, \varphi\rangle| \leq c\, P_{\mathsf{K},j}(\varphi), \quad \text{for any } \varphi \in \mathcal{D}(\Omega) \text{ with } \operatorname{supp}\varphi \subset \mathsf{K}.$$

Consequently, from (13.12) we have

$$|\langle \mathscr{D}^{\boldsymbol{\alpha}}u,\ \phi\rangle| = |\langle u,\ \partial^{\boldsymbol{\alpha}}\phi\rangle| \leq cP_{\mathsf{K},j}\left(\partial^{\boldsymbol{\alpha}}\phi\right) = cP_{\mathsf{K},j+|\boldsymbol{\alpha}|}\left(\phi\right),$$

which, by Definition 13.8, shows that $\mathscr{D}^{\boldsymbol{\alpha}}u$ is a distribution.

Definition 13.13 (equal in the distributional sense). Let u and v be two distributions in $\mathcal{D}'(\Omega)$. We say that u and v are equal in the distributional sense iff

$$\langle u, \phi\rangle = \langle v, \phi\rangle, \quad \forall \phi \in \mathcal{D}(\Omega).$$

In the next two examples, we show (in fact identify) that $\mathscr{D}^{\boldsymbol{\alpha}} = \partial^{\boldsymbol{\alpha}}$ for distributions that possess $|\boldsymbol{\alpha}|$-order classical derivatives. This confirms that the distributional derivative notion defined in (13.12) is indeed a generalization of the classical derivative concept.

[4] There is no universal notation for the adjoint operator $\mathscr{A}^*$ or the transport operator $\mathscr{A}'$. In fact, some use $\mathscr{A}'$ for the adjoint and $\mathscr{A}^*$ for the transpose.

Example 13.4. Consider $u \in \mathcal{C}^2(\Omega) \subset \mathbb{L}^1_{loc}(\Omega)$ with $\Omega \subset \mathbb{R}$. The regular distribution U (see (13.7)) induced by u becomes

$$\langle U,\ \phi\rangle := \int_\Omega u(x)\ \phi(x)\ dx = (u,\phi)_{\mathbb{L}^2(\Omega)}\,, \quad \forall \phi \in \mathcal{D}(\Omega)\,, \tag{13.13}$$

We are interested in knowing what the second order distributional derivative $\mathscr{D}^2 U$ of the regular distribution U looks like. For this example, $\partial^2 u$ is the second-order total derivative of u, and from Corollary 13.2 we know that $\partial^2 u \in \mathbb{L}^1_{loc}(\Omega)$, and thus itself generates a regular distribution denoted by V. Starting from the definition (13.12) for $|\alpha| = \alpha = 2$ and integrating by parts twice give

$$\left\langle (-1)^2 \mathscr{D}^2 U, \phi \right\rangle \overset{(13.12)}{=\!=} \langle U, \partial^2 \phi\rangle \overset{(13.13)}{=\!=} \left(u, \frac{d^2\phi}{dx^2}\right)_{\mathbb{L}^2(\Omega)}$$
$$\overset{\text{integration by parts twice}}{=\!=\!=} \left(\frac{d^2 u}{dx^2}, \phi\right)_{\mathbb{L}^2(\Omega)} \overset{\text{definition of } V}{=\!=} \langle V, \phi\rangle\,, \quad \forall \phi \in \mathcal{D}(\Omega)\,,$$

where the boundary terms from integration by parts vanish since ϕ has compact support in Ω. Therefore, by Lemma 13.1, we have

$$\mathscr{D}^2 U = V,$$

that is, the second-order distributional derivative of U—the regular distribution induced by u—is the regular distribution induced by $\dfrac{d^2 u}{dx^2}$. A better understanding is perhaps to recognize that both are regular distributions induced by $\dfrac{d^2 u}{dx^2}$, and thus are identical in the distributional sense. From the discussion in Remark 13.5, we identify the regular distribution $\mathscr{D}^2 U$ with its generator $\dfrac{d^2 u}{dx^2}$ in additional to identifying U with u, and in the distributional sense we write

$$\mathscr{D}^2 u = \frac{d^2 u}{dx^2},$$

that is, we say the distributional derivative reduces to the classical derivative. What such a (conventional) statement means is

$$\left\langle \mathscr{D}^2 u, \phi \right\rangle = \left(\frac{d^2 u}{dx^2}, \phi\right)_{\mathbb{L}^2(\Omega)}\,, \quad \forall \phi \in \mathcal{D}(\Omega)\,.$$

The identification becomes precise when we introduce the weak derivative in Definition 13.14.

Example 13.5 ($\mathscr{D}^{\boldsymbol{\alpha}}$ reduces to $\partial^{\boldsymbol{\alpha}}$). This example generalizes Example 13.4. Consider $\Omega \subseteq \mathbb{R}^n$, a multi-index $\boldsymbol{\alpha}$, $u \in \mathcal{C}^{\boldsymbol{\alpha}}(\Omega) \subset \mathbb{L}^1_{loc}(\Omega)$. The regular distribution (13.7) induced by u, identified with u, becomes

$$\langle u,\ \phi\rangle := \int_\Omega u(\boldsymbol{x})\ \phi(\boldsymbol{x})\ d\boldsymbol{x} = (u,\phi)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega).$$

For any $\phi \in \mathcal{D}(\Omega)$, we have

$$\langle (-1)^{\boldsymbol{\alpha}} \mathscr{D}^{\boldsymbol{\alpha}} u, \phi\rangle = \langle u, \partial^{\boldsymbol{\alpha}}\phi\rangle = (u, \partial^{\boldsymbol{\alpha}}\phi)_{\mathbb{L}^2(\Omega)} = \left((-1)^{|\boldsymbol{\alpha}|}\, \partial^{\boldsymbol{\alpha}} u, \phi\right)_{\mathbb{L}^2(\Omega)},$$

where we have used the integration by parts formula (13.5) in the last equality. Thus, for $|\boldsymbol{\alpha}|$-order differentiable functions, not only

$$\mathscr{D}^{\boldsymbol{\alpha}} = \partial^{\boldsymbol{\alpha}},$$

in the distributional sense due to Definition 13.13 but also $\mathscr{D}^{\boldsymbol{\alpha}}$ is the regular distribution induced by $\partial^{\boldsymbol{\alpha}}$ (see Definition 13.10). The identification becomes precise when we introduce the weak derivative in Definition 13.14.

Example 13.6 (Dirac delta is the distributional derivative of the Heaviside function). Let us continue with Example 13.3 and now we are interested in finding the distribution derivative $\mathscr{D}f$ of f. Clearly, the classical derivative for f does not exist (at least at $x = 0$) as f is a discontinuous function. The distributional derivative of[5] f is, however, well-defined as we now show. From definition (13.12), we have

$$\begin{aligned} -\langle \mathscr{D}f, \phi\rangle = \langle f, \partial\phi\rangle = (f, \partial\phi)_{\mathbb{L}^2(\mathbb{R})} &= \int_0^\infty \partial\phi(x)\ dx = \int_0^\infty \frac{d\phi}{dx}\, dx \\ &= \phi(\infty) - \phi(0) = -\phi(0) = -\langle \delta, \phi\rangle, \quad \forall \phi \in \mathcal{D}(\mathbb{R}), \end{aligned}$$

where we have used Definition 13.11 for the Dirac delta in the last inequality. By Definition 13.13, we conclude

$$\mathscr{D}f = \delta$$

in the distributional sense. We can continue this procedure to compute any higher-order distributional derivatives for f to obtain

$$\mathscr{D}^m f = \mathscr{D}^{m-1}\delta,$$

in the distributional sense. For example, $\forall \phi \in \mathcal{D}(\mathbb{R})$:

$$\left\langle \mathscr{D}^2 f, \phi\right\rangle = \left\langle f, \partial^2\phi\right\rangle = \int_0^\infty \frac{\partial^2\phi}{\partial x^2}\, dx = \frac{d\phi}{dx}(0) = \langle \delta, \partial\phi\rangle = \langle \mathscr{D}\delta, \phi\rangle,$$

where we have used the distributional derivative definition (13.12) in the first and the last equalities.

[5] of the regular distribution induced by f to be precise, but again from Remark 13.5, we do not distinguish them.

Definition 13.14 (Weak derivatives). Let $f \in \mathbb{L}^1_{loc}(\Omega)$ with $\Omega \subseteq \mathbb{R}^n$ and a multi-index $\boldsymbol{\alpha}$. If there exists a locally integrable function $g \in \mathbb{L}^1_{loc}(\Omega)$ such that:

$$\left\langle (-1)^{|\boldsymbol{\alpha}|} \mathscr{D}^{\boldsymbol{\alpha}} f,\ \phi \right\rangle = \langle f,\ \partial^{\boldsymbol{\alpha}} \phi \rangle = \left((-1)^{|\boldsymbol{\alpha}|} g, \phi \right)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega),$$

then, g is called as the weak derivative of f and we write $\mathscr{D}^{\boldsymbol{\alpha}} f = g$ (in distributional sense). In other words, a weak derivative of a distribution, if exists, is a locally integrable function.[6]

Note that a weak derivative is a special case of the distributional derivative. From Example 13.5, the classical derivatives are in turn special cases of weak derivatives.

Lemma 13.2 (Uniqueness of weak derivative). *Consider $f \in \mathbb{L}^1_{loc}(\Omega)$ and a multi-index $\boldsymbol{\alpha}$. Let $g,\ \tilde{g} \in \mathbb{L}^1_{loc}(\Omega)$ be an $|\boldsymbol{\alpha}|$-order weak derivative of f, that is,*

$$\left\langle (-1)^{|\boldsymbol{\alpha}|} \mathscr{D}^{\boldsymbol{\alpha}} f,\ \phi \right\rangle = \left((-1)^{|\boldsymbol{\alpha}|} g, \phi \right)_{\mathbb{L}^2(\Omega)} = \left((-1)^{|\boldsymbol{\alpha}|} \tilde{g}, \phi \right)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega),$$

then, $g = \tilde{g}$ for almost everywhere.

Proof. The proof is obvious due to Lemma 13.1.

Example 13.7. Let us now revisit Example 13.4. From

$$\left\langle (-1)^2 \mathscr{D}^2 U, \phi \right\rangle = \left(\frac{d^2 u}{dx^2}, \phi \right)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega),$$

we conclude that $\mathscr{D}^2 U = \dfrac{d^2 u}{dx^2}$ is the second-order weak derivative of u. That is the weak derivatives up to the second order of u coincide with the classical derivatives.

Similarly, let us revisit Example 13.5, and we see that the weak derivatives up to order $|\boldsymbol{\alpha}|$ not only exist but they are identical to the corresponding classical derivatives. Again, this confirms that the generalized derivative, here weak derivative, is consistent in the sense that when the function under consideration possesses classical derivatives, the classical and generalized derivatives coincide.

Example 13.8. Consider the following discontinuous function (and thus not differentiable in the classical sense):

$$f(x) = \begin{cases} 0, & \text{for x rational,} \\ 2 + \sin x, & \text{for x irrational.} \end{cases}$$

[6] In fact they are the same almost everywhere, i.e. they are in the same equivalent class, and we do not distinguish them.

Clearly, f is locally integrable on $\Omega = \mathbb{R}$. Now computing its distributional derivative, we have

$$\langle \mathscr{D} f,\ \phi \rangle = -\left\langle f,\ \frac{d\phi}{dx} \right\rangle = -\int_{-\infty}^{\infty} (2 + \sin x) \frac{d\phi}{dx}\, dx$$
$$= \int_{-\infty}^{\infty} \cos x\ \phi\ dx = (\cos,\ \phi)_{\mathbb{L}^2(\mathbb{R})}\,.$$

Since cos is locally integrable on Ω, by Definition 13.14 cos is the weak derivative of f. Following the same procedure, it is easy to see that f possesses weak derivatives of all orders. In particular, the nth-order weak derivative of f is given by

$$\mathscr{D}^n f = (-1)^{n-1} \begin{cases} \cos(x) & \text{if } n \text{ is odd} \\ \sin(x) & \text{otherwise} \end{cases}, \quad \forall n \geq 1.$$

Example 13.9. Consider the distributional derivatives of the Heaviside function in Example 13.6, we see that the Heaviside function does not have weak derivatives as δ is not locally integrable (see (13.10)).

13.1.5 Sobolev spaces and some applications

We limit ourselves to a brief introduction to $\mathbb{L}^2$-based Sobolev spaces of integer order. (Extensive presentations on Sobolev spaces can be consulted from [3, 102] and references therein). The goal of this section is to use the above developments on weak derivatives to introduce Sobolev spaces that are needed to properly understand/derive the Green functions in section 13.2 and the PDE examples in Chapter 15. Similar to other sections, we consider $\Omega \subseteq \mathbb{R}^n$, for some integer n.

Definition 13.15 (Sobolev spaces $\mathbb{H}^m(\Omega)$). Let m be a nonnegative integer and $\boldsymbol{\alpha}$ be a multi-index. We define $\mathbb{H}^m(\Omega)$ as:

$$\mathbb{H}^m(\Omega) = \left\{ u \in \mathbb{L}^2(\Omega) :\ \mathscr{D}^{\boldsymbol{\alpha}} u \in \mathbb{L}^2(\Omega),\ \forall\, |\boldsymbol{\alpha}| \leq m \right\},$$

that is, $\mathbb{H}^m(\Omega)$ is the collection of all functions whose distributional derivatives up to order m are square-integrable ($\mathbb{L}^2$-functions). Clearly, $\mathbb{H}^0(\Omega) = \mathbb{L}^2(\Omega)$ and $\mathbb{H}^p(\Omega) \subset \mathbb{H}^m(\Omega)$ for $p \geq m$.

The following result is well-known [3, 102, 11].

Lemma 13.3. *For any non-negative integer m, $\mathbb{H}^m(\Omega)$ is a Hilbert space with the following inner product*

$$(f,g)_{\mathbb{H}^m(\Omega)} := \sum_{|\boldsymbol{\alpha}|\leq m} (\mathscr{D}^{\boldsymbol{\alpha}} f, \mathscr{D}^{\boldsymbol{\alpha}} g)_{\mathbb{L}^2(\Omega)}, \quad \forall f,g \in \mathbb{H}^m(\Omega), \tag{13.14}$$

and the induced norm

$$\|f\|_{\mathbb{H}^m(\Omega)} = \sqrt{(f,f)_{\mathbb{H}^m(\Omega)}}, \quad \forall f \in \mathbb{H}^m(\Omega). \tag{13.15}$$

Definition 13.16 ($\mathbb{H}_0^m(\Omega)$). For a non-negative integer m, we define $\mathbb{H}_0^m(\Omega)$ as the closure of $\mathcal{D}(\Omega)$ under the $\mathbb{H}^m$-norm (13.15):

$$\mathbb{H}_0^m(\Omega) = \overline{\mathcal{D}(\Omega)}^{\mathbb{H}^m}.$$

We present two consequences of Definition 13.16.

Corollary 13.4. *$\mathbb{H}_0^m(\Omega)$, equipped with the $\mathbb{H}^m$-inner product (13.14), is a Hilbert space.*

Corollary 13.5 (Integration by parts for $\mathbb{H}^m(\Omega)$). *Let m be any non-negative integer. For any $f \in \mathbb{H}^m(\Omega)$ and any $g \in \mathbb{H}_0^m(\Omega)$, the following integration by parts formula holds:*

$$\left((-1)^{|\boldsymbol{\alpha}|}\mathscr{D}^{\boldsymbol{\alpha}} f,\ g\right)_{\mathbb{L}^2(\Omega)} = (f,\ \mathscr{D}^{\boldsymbol{\alpha}} g)_{\mathbb{L}^2(\Omega)}, \quad \forall\, |\boldsymbol{\alpha}| \leq m. \tag{13.16}$$

Proof. From Corollary 13.2 and Definition 13.14, we have

$$\left((-1)^{|\boldsymbol{\alpha}|}\mathscr{D}^{\boldsymbol{\alpha}} f,\ \phi\right)_{\mathbb{L}^2(\Omega)} = (f,\ \partial^{\boldsymbol{\alpha}}\phi)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega), \forall\, |\boldsymbol{\alpha}| \leq m. \tag{13.17}$$

By Definition 13.16, there exists a sequence $\{\varphi_n\}_{n=1}^{\infty} \subset \mathcal{D}(\Omega)$ converges to g in the $\mathbb{H}^m$-norm. In particular, $\|\partial^{\boldsymbol{\alpha}}\varphi_n - \mathscr{D}^{\boldsymbol{\alpha}} g\|_{\mathbb{L}^2(\Omega)} \to 0$, for any $|\boldsymbol{\alpha}| \leq m$, as n approaches ∞. Replacing ϕ with φ_n in (13.17), using the fact that the $\mathbb{L}^2$-inner product is continuous in both of its arguments in the $\mathbb{L}^2$-norm, and passing to the limit conclude the proof.

Remark 13.7. Note that the integration by parts formula for $\mathbb{H}^m$ in (13.16) is similar to the standard integration by parts for differentiable functions except without having the boundary terms. For open and bounded domain $\Omega \subset \mathbb{R}^n$, this suggests that all the (directional normal) derivatives of g up to order $m-1$ vanish on $\partial\Omega$ since f and any of its weak derivatives do not vanish on $\partial\Omega$ in general. A rigorous justification of this requires the study of the trace of functions in $\mathbb{H}^m$ (see, e.g., [132, 102]), but this is out of the scope of the book.

Definition 13.17 (Topological dual of $\mathbb{H}_0^m(\Omega)$). We denote by $\mathbb{H}^{-m}(\Omega)$ the space of continuous linear functionals on $\mathbb{H}_0^m(\Omega)$ with the operator norm (see Theorem 5.1)

$$\|u\|_{\mathbb{H}^{-m}} = \sup_{\varphi \in \mathbb{H}_0^m(\Omega)} \frac{\langle u, \varphi \rangle_{\mathbb{H}^m}}{\|\varphi\|_{\mathbb{H}^m}}$$

Proposition 13.3 (Characterization of $\mathbb{H}^{-m}(\Omega)$). *There holds:*

$$\mathbb{H}^{-m}(\Omega) := span\left\{ \mathscr{D}^{\boldsymbol{\alpha}} f : |\boldsymbol{\alpha}| \le m \text{ and } f \in \mathbb{L}^2(\Omega) \right\}.$$

Proof. Suppose $f \in \mathbb{L}^2(\Omega)$ and $\boldsymbol{\alpha}$ is a multi-index such that $|\boldsymbol{\alpha}| \le m$. By Definition 13.12, $\forall \phi \in \mathcal{D}(\Omega)$ we have

$$\begin{aligned} \langle \mathscr{D}^{\alpha} f, \phi \rangle = (-1)^{|\boldsymbol{\alpha}|} \langle f, \partial^{\boldsymbol{\alpha}} \phi \rangle &= (-1)^{|\boldsymbol{\alpha}|} (f, \partial^{\boldsymbol{\alpha}} \phi)_{\mathbb{L}^2(\Omega)} \\ &\le \|f\|_{\mathbb{L}^2(\Omega)} \|\partial^{\boldsymbol{\alpha}} \phi\|_{\mathbb{L}^2(\Omega)} \le \|f\|_{\mathbb{L}^2(\Omega)} \|\phi\|_{\mathbb{H}^m(\Omega)}, \end{aligned}$$

where have used Corollary 13.2 in the second equality and (13.14) in the last inequality. That is, the disbution $\mathscr{D}^{\alpha} f$ is linear and continuous on $\mathcal{D}(\Omega)$ in the $\mathbb{H}^m$-norm. By the extension principle Lemma 12.1, $\mathscr{D}^{\alpha} f$ can be extended uniquely to a linear and continuous functional on $\mathbb{H}_0^m(\Omega)$, again denoted by $\mathscr{D}^{\alpha} f$. That is, $\mathscr{D}^{\alpha} f \in \mathbb{H}^{-m}(\Omega)$.

Conversely, suppose $u \in \mathbb{H}^{-m}(\Omega)$. Since $\mathbb{H}_0^m(\Omega)$ is a Hilbert space (see Corollary 13.4), by the Riesz representation Theorem 5.1 there exists a unique $g \in \mathbb{H}_0^m(\Omega)$ such that

$$u(v) := \langle u, v \rangle_{\mathbb{H}_0^m(\Omega)} = (g, v)_{\mathbb{H}_0^m(\Omega)}, \quad \forall v \in \mathbb{H}_0^m(\Omega),$$

where we recall Remark 5.4 that the notation $\langle u, v \rangle_{\mathbb{H}_0^m(\Omega)}$ means u and v is a duality pairing in $\mathbb{H}_0^m(\Omega)$. It follows from Definition 13.16 that

$$u(\phi) = (g, \phi)_{\mathbb{H}_0^m(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega).$$

which defines the distribution u via $\mathbb{H}^m$-inner product (compared to regular distributions defined via $\mathbb{L}^2$-inner product in (13.7)). In other words, g defines/induces the distribution u via its $\mathbb{H}^m$-inner product with test functions. Using distributional derivation Definition 13.12 gives

$$u(\phi) = \left\langle \sum_{|\boldsymbol{\alpha}| \le m} (-1)^{|\boldsymbol{\alpha}|} \mathscr{D}^{\boldsymbol{\alpha}} g, \phi \right\rangle, \quad \forall \phi \in \mathcal{D}(\Omega),$$

that is, in the distributional sense we have the following identification:

$$u = \sum_{|\boldsymbol{\alpha}| \le m} (-1)^{|\boldsymbol{\alpha}|} \mathscr{D}^{\boldsymbol{\alpha}} (\mathscr{D}^{\boldsymbol{\alpha}} g), \tag{13.18}$$

which completes the proof as by Definition 13.15 $\mathscr{D}^{\boldsymbol{\alpha}} g \in \mathbb{L}^2(\Omega)$ for all $\boldsymbol{\alpha}$ with $|\boldsymbol{\alpha}| \leq m$.

Corollary 13.6 ($\delta \in \mathbb{H}^{-1}(\Omega)$). *Let us define the n-dimensional Dirac delta distribution in Definition 13.11 as the tensor product of one-dimensional Dirac delta functions. Then $\delta \in \mathbb{H}^{-1}(\Omega)$. In particular, there exists a function $g \in \mathbb{H}_0^1(\Omega)$ such that*

$$\delta = g - \sum_{|\boldsymbol{\alpha}|=1} \mathscr{D}^{\boldsymbol{\alpha}} \left(\mathscr{D}^{\boldsymbol{\alpha}} g \right) \tag{13.19}$$

in the distributional sense.

Proof. The proof is obvious due to the tensor product structure of δ, Example 13.6, and Proposition 13.3.

Remark 13.8. Note that in the representation (13.18) of distributions in $\mathbb{H}^{-m}$, and in particular of the Dirac delta in (13.19), the inner distributional derivative $\mathscr{D}^{\boldsymbol{\alpha}}$ is a weak derivative since $g \in \mathbb{H}_0^m(\Omega)$ while the outer is a distribution derivative in general. The Riesz representation Theorem 5.1 tells us that there is a unique represented ℓ in the Hilbert space $\mathbb{U}$ for a given linear continuous functional $\mathscr{L}$, but it does not tell us how to find ℓ in general. When $\mathbb{U} = \mathbb{H}_0^m$, Proposition 13.3 tells us that if we consider the distributional differential equation (13.18) given $u \in \mathbb{H}^{-m}$, we know that its unique solution is the Riesz representer g of u.

The next example is a nice exercise on the interplay among distributional derivatives, weak derivatives, dual spaces, and dense subsets to study the relationship between Sobolev spaces and their topological duals.

Example 13.10. Let $u \in \mathbb{H}^m(\Omega)$ where m is an integer. Define $v := \mathscr{D}^{\boldsymbol{\gamma}} u$, where $\boldsymbol{\gamma}$ is a multi-index. Clearly, if m is a non-negative integer, $|\boldsymbol{\gamma}| \leq m$, and u resides in $\mathcal{C}^m(\Omega)$, then we know that $v \in \mathcal{C}^{m-|\boldsymbol{\gamma}|}(\Omega)$. Consequently, if the distributional/weak derivatives are generalizations of the classical derivatives, we expect that v resides in $\mathbb{H}^{m-|\boldsymbol{\gamma}|}(\Omega)$. The objective of this example is to rigorously confirm this hypothesis for an arbitrary integer m.

Step 1. First, let us consider non-negative m. By Definition 13.15 of Sobolev spaces, $\mathscr{D}^{\boldsymbol{\alpha}} u \in \mathbb{L}^2(\Omega)$ for all $|\boldsymbol{\alpha}| \leq m$. From distributional derivative Definition 13.12, if a multi-index $\boldsymbol{\beta}$ is such that $|\boldsymbol{\gamma} + \boldsymbol{\beta}| = |\boldsymbol{\gamma}| + |\boldsymbol{\beta}| \leq m$ we have

$$\begin{aligned} &\left\langle (-1)^{|\boldsymbol{\beta}|} \mathscr{D}^{\boldsymbol{\beta}} v, \phi \right\rangle = \left\langle v, \partial^{\boldsymbol{\beta}} \phi \right\rangle = \left\langle \mathscr{D}^{\boldsymbol{\gamma}} u, \partial^{\boldsymbol{\beta}} \phi \right\rangle \\ &= \left\langle (-1)^{|\boldsymbol{\beta}|} \mathscr{D}^{\boldsymbol{\gamma}+\boldsymbol{\beta}} u, \phi \right\rangle = \left((-1)^{|\boldsymbol{\beta}|} \mathscr{D}^{\boldsymbol{\gamma}+\boldsymbol{\beta}} u, \phi \right)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega). \end{aligned}$$

In other words, by Definition 13.14, v possesses weak derivatives up to order $|\boldsymbol{\beta}|$ as long as $|\boldsymbol{\beta}| \leq m - |\boldsymbol{\gamma}|$. Thus, by Definition 13.15, $v \in \mathbb{H}^{m-|\boldsymbol{\gamma}|}(\Omega)$.

Step 2. Now for $m = -q \leq 0$, we have

$$\langle v, \phi \rangle = (-1)^{|\boldsymbol{\gamma}|} \langle u, \mathscr{D}^{\boldsymbol{\gamma}} \phi \rangle , \quad \forall \phi \in \mathcal{D}(\Omega) .$$

Since, from Definition 13.16, $\mathcal{D}(\Omega)$ is dense in $\mathbb{H}_0^p(\Omega)$ in the $\mathbb{H}^p$-norm for any non-negative integer p, we can replace ϕ by w (see Lemma 12.1)

$$\langle v, w \rangle = (-1)^{|\boldsymbol{\gamma}|} \langle u, \mathscr{D}^{\boldsymbol{\gamma}} w \rangle , \quad \forall w \in \mathbb{H}_0^p(\Omega) ,$$

as long as $p - |\boldsymbol{\gamma}| = q$, because then $\mathscr{D}^{\boldsymbol{\gamma}} w \in \mathbb{H}^q$ by the Step 1, and thus the right-hand side is a valid duality pairing for $\mathbb{H}^{-q}(\Omega)$ and $\mathbb{H}_0^q(\Omega)$. Consequently,

$$|\langle v, w \rangle| \leq \|u\|_{\mathbb{H}^{-q}(\Omega)} \|w\|_{\mathbb{H}^q(\Omega)} \leq \|u\|_{\mathbb{H}^{-q}(\Omega)} \|w\|_{\mathbb{H}^{|\boldsymbol{\gamma}|+q}(\Omega)}$$

which implies that $v \in \mathbb{H}^{-q-|\boldsymbol{\gamma}|}(\Omega) = \mathbb{H}^{m-|\boldsymbol{\gamma}|}(\Omega)$.

We are in the position to revisit the motivational example in subsection 13.1.1 with two objectives in mind. First, we shall show that the classical setting is a particular case of a generalized setting with weak derivative, that is, a solution to the classical setting is also a solution to the weak setting. Second, we show that while the former does not make sense when f is rough (discontinuous), not only is the latter meaningful, but it also admits a solution. Finally we show that when f is sufficiently in the weak setting, the weak solution is classical solution of the classical setting (13.20), and this completes our demonstration that the weak setting is indeed a generalization of the classical setting.

Example 13.11. Back to the motivational example in subsection 13.1.1. With $\Omega = (0,1)$, if $f \in \mathbb{L}^2(\Omega)$ with real value, it is locally integrable owing to Corollary 13.2. Thus, the classical setting

$$\frac{du}{dx} = f, \tag{13.20}$$

does not have a classical solution if $f \in \mathbb{L}^2(\Omega) \setminus \mathcal{C}(\Omega)$. Suppose $f \in \mathcal{C}(\Omega)$ so that (13.20) has a classical solution $u \in \mathcal{C}^1(\Omega)$. For all $\phi \in \mathcal{D}(\Omega)$, we have

$$(f, \phi)_{\mathbb{L}^2(\Omega)} = \left(\frac{du}{dx}, \phi \right)_{\mathbb{L}^2(\Omega)} = - \left(u, \frac{d\phi}{dx} \right)_{\mathbb{L}^2(\Omega)} = - (u, \partial \phi)_{\mathbb{L}^2(\Omega)} . \tag{13.21}$$

Using the fact that u is differentiable and Example 13.7, we can write (13.21) equivalently as

$$\mathscr{D} u = f, \tag{13.22}$$

where $\mathscr{D} u$ is the weak derivative of u. *Thus, we have shown that a classical solution of the classical setting* (13.20) *is also a solution to ODE in the weak form* (13.22).

The interesting point of weak form (13.22) is that it is also valid for $f \in \mathbb{L}^2(\Omega)$, as opposed to classical setting (13.20) which admits only classical solutions for sufficiently smooth f. In other words, (13.22) allows us to look for solutions in bigger spaces when f is "rough" (such as discontinuous functions). The hope is that a sufficiently big space may contain a solution. A reasonable general setting for (13.22) is therefore to look for a weak solution $u \in \mathbb{H}^1(\Omega)$. The existence and uniqueness of the solution will be shown in Example 13.12. *It is important to point out that such a weak solution u may not be a solution of the classical setting (13.20) as it is not classically differentiable in general.*

We have shown that we can go from classical setting (13.20) to weak setting (13.22) if f is such that a solution u is differentiable. The question is: could we show that the weak setting (13.22) reduces to classical setting (13.20) when $f \in \mathcal{C}(\Omega)$? Only *when the answer is affirmative, is (13.22) truly a generalization of (13.20)*. The answer is indeed quite straightforward. Since $u \in \mathbb{H}^1(\Omega)$, u and $\mathscr{D}u$ reside in $\mathbb{L}^2(\Omega)$ by Definition 13.15. Now since f is continuous, $\mathscr{D}u = f$ is continuous.[7] By invoking the regularity result in [28, Theorem 8.2 and Remark 6], we conclude that $u \in \mathcal{C}^1(\Omega)$. Thus, the weak solution is indeed classical and it is a solution of the classical setting (13.20) when $f \in \mathcal{C}(\Omega)$.

Example 13.12 (Characterization of $\mathbb{H}_0^1(0,1)$). We continue Example 13.11 and we aim to show the existence and uniqueness of the solution in (13.22) (achieving the same results using the Banach-Nečas-Babuška theorem will be provided in Example 15.2). To that end, let us consider the following closed and dense subspace of $\mathbb{H}^1[0,1]$, and thus still Hilbert,

$$\mathbb{H}_0^1[0,1] := \left\{ g \in \mathbb{H}^1[0,1] : g(0) = 0 \right\}.$$

Solving for a solution of (13.22) is the same as asking for the characterization of functions in $\mathbb{H}_0^1[0,1]$, that is, what a function looks like if itself and its first order weak derivative are square integrable. From the fundamental theorem of calculus for $\mathbb{C}^1[0,1]$, we expect that $u(x)$ is given in the form

$$\tilde{u}(x) = \int_0^x \mathscr{D}u(t)\, dt. \tag{13.23}$$

We now make this rigorous. First, the right-hand side of (13.23) is meaningful by the Cauchy-Schwarz inequality (13.4):

$$\left| \int_0^x \mathscr{D}u(t)\, dt \right| \leq \sqrt{x}\, \|\mathscr{D}u\|_{\mathbb{L}^2} < \infty. \tag{13.24}$$

Second, $\tilde{u}$ is a continuous function. Indeed, let $x_n \xrightarrow{n\to\infty} x$ and we have

[7] Technically, $\mathscr{D}u$ is equal to f almost everywhere. We simply pick $\mathscr{D}u = f$ pointwise as both belong to the same equivalent class.

$$\lim_{n\to\infty}\left[\tilde{u}\left(x\right)-\tilde{u}\left(x_n\right)\right]=\lim_{n\to\infty}\int_0^1 \mathbf{1}_{[x_n,x]}\left(t\right)\mathscr{D}u\left(t\right)\,dt$$
$$=\int_0^1 \lim_{n\to\infty}\mathbf{1}_{[x_n,x]}\left(t\right)\mathscr{D}u\left(t\right)\,dt=0,$$

where, without loss of generality, we assume that x_n approaches x from below and thus the indicator function $\mathbf{1}_{[x_n,x]}\left(t\right)$ makes sense, and we have invoked the dominated convergence theorem[8] in the second last equality.

Third, the weak derivative of $\tilde{u}$ is $\mathscr{D}u$. Indeed, let us denote $\Omega=(0,1)$, by Corollary 13.2 we have

$$-\left\langle\mathscr{D}\tilde{u},\phi\right\rangle=\left(\tilde{u},\partial\phi\right)_{\mathbb{L}^2}=\int_0^1\partial\phi\left(x\right)\int_0^x\mathscr{D}u\left(t\right)\,dt\,dx$$
$$=\int_0^1\mathscr{D}u\left(t\right)\int_t^1\partial\phi\left(x\right)\,dx\,dt=-\int_0^1\mathscr{D}u\left(t\right)\phi\left(t\right)\,dt,\quad\forall\phi\in\mathcal{D}\left(\Omega\right),$$

and thus $\mathscr{D}\tilde{u}=\mathscr{D}u$ by Definition 13.14.

Fourth, $\tilde{u}=u$ almost everywhere, that is, two $\mathbb{L}^2$ functions with the same weak derivative must be in the same equivalent class. Indeed, from $\mathscr{D}\tilde{u}=\mathscr{D}u$ we conclude that $\tilde{u}-u$ must be constant almost everywhere (see, e.g., [28, Lemma 8.1] and [53, Lemma B.29]). Since both $\tilde{u}$ and u vanish at $x=0$, $\tilde{u}=u$ almost everywhere and $\tilde{u}$ is the unique continuous representation of u. This is consistent with the Sobolev embedding theorem [11, 28, 3, 102, 132] which shows that $\mathbb{H}_0^1\left[0,1\right]$ is embedded in the space of continuous functions. Clearly, embedding means that each function in $\mathbb{H}_0^1\left[0,1\right]$ is equal almost everywhere to its unique continuous representative in the same equivalent class.

We conclude that $u\in\mathbb{H}_0^1\left[0,1\right]$ iff

- The weak derivative of u is square integrable, i.e. $\mathscr{D}u\in\mathbb{L}^2\left[0,1\right]$, and
- u is continuous and obeys the fundamental theorem of calculus

$$u\left(x\right)=\int_0^x\mathscr{D}u\left(t\right)\,dt.$$

It follows that there is a unique solution in $\mathbb{H}_0^1\left(0,1\right)$ to (13.22).

Similar to Example 13.11, we now discusses the classical and weak settings for an elliptic partial differential equation (PDE).

Example 13.13. Let $\Omega\subset\mathbb{R}^n$ be open and bounded. Consider the following elliptic PDE:

$$\begin{aligned}-\Delta u+u&=f,\quad\text{in }\Omega,\\ u&=0,\quad\text{in }\partial\Omega,\end{aligned}\tag{13.25}$$

[8] It is obvious that $\mathbf{1}_{[x_n,x]}\left(t\right)\mathscr{D}u\left(t\right)\xrightarrow{n\to\infty}0$ almost everywhere, $\left|\mathbf{1}_{[x_n,x]}\left(t\right)\mathscr{D}u\left(t\right)\right|\le\left|\mathscr{D}u\left(t\right)\right|$. Together with the fact that $\left|\mathscr{D}u\left(t\right)\right|$ is integrable by (13.24), we can apply the dominated convergence theorem to switch the limit and the integration.

where $\Delta(\cdot) := \sum_{i=1}^{n} \partial_{\boldsymbol{x}_i}(\partial_{\boldsymbol{x}_i}(\cdot))$ is the Laplacian.

If $f \in \mathcal{C}(\Omega)$, then the classical setting (13.25) in which we look for a solution $u \in \mathcal{C}^2(\Omega)$ makes sense. However, when f is discontinuous, such as $f \in \mathbb{L}^2(\Omega) \setminus \mathcal{C}(\Omega)$, the classical setting does not make sense, that is, there is no classical solution.

Now, suppose, for the moment, there is a solution $u \in C^2(\Omega)$ (with f necessarily being in $\mathcal{C}(\Omega)$ of course), then from (13.25) we have

$$(-\Delta u + u, \phi)_{\mathbb{L}^2(\Omega)} = (f, \phi)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega),$$

which, after integration by parts, becomes

$$(\nabla u, \nabla \phi)_{\mathbb{L}^2(\Omega)} + (u, \phi)_{\mathbb{L}^2(\Omega)} = (f, \phi)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega), \tag{13.26}$$

which, by using the definition of weak derivative in Definition 13.14, can be equivalently written as

$$-\Delta^{\mathscr{D}} u + u = f, \tag{13.27}$$

where we have used $\Delta^{\mathscr{D}}(\cdot) := \sum_{i=1}^{n} \mathscr{D}_{\boldsymbol{x}_i}(\mathscr{D}_{\boldsymbol{x}_i}(\cdot))$ to denote the weak Laplacian, which is nothing more than the classical Laplacian (recall Example 13.7). Thus, we have shown that if u is a solution for the classical setting (13.25) it is also a solution for the weak setting (13.27).

Now suppose that $f \in \mathbb{L}^2(\Omega) \setminus \mathcal{C}(\Omega)$. In this case, the classical setting (13.25) is not meaningful and there is no classical solution. The question is if the weak setting (13.27) has a solution? From crefeq:weakEllipticPDE, a natural general setting is to search for a solution in a bigger space of $\mathbb{H}^1(\Omega)$ that contains functions whose weak derivatives up to the first order are in $\mathbb{L}^2(\Omega)$ (see Definition 13.15). If we do that, using the definition of the $\mathbb{H}^1$ inner product in (13.14), we can write (13.26) as

$$(u, \phi)_{\mathbb{H}^1(\Omega)} = (f, \phi)_{\mathbb{L}^2(\Omega)}, \quad \forall \phi \in \mathcal{D}(\Omega). \tag{13.28}$$

Clearly, by Example 13.7, a solution to the classical setting (13.25) is also a solution of the $\mathbb{H}^1$-setting (13.28). However, not only does the latter make sense for any $f \in \mathbb{L}^2(\Omega)$ (discontinuous or not), but it also has a unique solution as we now show. This should not be a surprise as we search for a solution in the bigger space $\mathbb{H}^1(\Omega)$, there is hope that we can find a solution.

Owing to Definition 13.16, $\mathcal{D}(\Omega)$ is dense in $\mathbb{H}_0^1(\Omega)$ with respect to the $\mathbb{H}^1$-norm, using a similar procedure as in the proof of Corollary 13.5, we see that a solution to (13.28) is a also a solution to the following equation (and vice versa):

$$(u, v)_{\mathbb{H}^1(\Omega)} = (f, v)_{\mathbb{L}^2(\Omega)}, \quad \forall v \in \mathbb{H}_0^1(\Omega). \tag{13.29}$$

Now, owing to Cauchy-Schwarz inequality (13.4) and $\mathbb{H}^1$-norm definition in (13.15), we have

$$\left|(f,v)_{\mathbb{L}^2(\Omega)}\right| \leq \|f\|_{\mathbb{L}^2(\Omega)} \|v\|_{\mathbb{L}^2(\Omega)} \leq \|f\|_{\mathbb{L}^2(\Omega)} \|v\|_{\mathbb{H}^1(\Omega)},$$

and thus f is linear and continuous on $\mathbb{H}_0^1(\Omega)$, i.e., $f \in \left[\mathbb{H}_0^1(\Omega)\right]^*$ (recall Definition 5.5 for the dual spaces). By the Riesz representation Theorem 5.1, there is a unique $\tilde{u} \in \mathbb{H}_0^1(\Omega)$ such that

$$(\tilde{u},v)_{\mathbb{H}^1(\Omega)} = (f,v)_{\mathbb{L}^2(\Omega)}.$$

Thus, (13.29) can be written equivalently as

$$(u-\tilde{u},v)_{\mathbb{H}^1(\Omega)} = 0, \quad \forall v \in \mathbb{H}_0^1(\Omega),$$

which implies that

$$u = \tilde{u}.$$

Thus, given an $f \in \mathbb{L}^2(\Omega)$, there is a unique solution $u \in \mathbb{H}_0^1(\Omega)$ for (13.29). *It is important to point out that such a solution $\tilde{u}$ may not be a solution of the classical setting (13.25) as it is not classically differentiable in general.*

Next, we discuss the generality of the weak form (13.29) (or equivalently (13.27)). In particular, if f is sufficiently smooth, then a weak solution is classical and it is a solution of the classical setting (13.25). As opposed to ODE (13.20), we need stringent conditions not only on f but also on the domain Ω in order to recover the classical solution from the weak form (13.29). In particular, if $f \in \mathbb{H}^m(\Omega)$ with[9] $m > n/2$ and for Ω with sufficiently smooth boundary $\partial\Omega$, then the unique solution $\tilde{u} \in \mathcal{C}^2\left(\overline{\Omega}\right)$ (see, e.g., [28, Theorem 9.25]).

13.2 Green functions as adjoint solutions

With the proper understanding of the Dirac delta function as a distribution and of distributional derivatives in section 13.1, we are now in the position to develop a proper understanding of the Green functions. The vital role of Green functions in understanding solutions of the corresponding PDE can be found in any textbooks on PDEs (see, e.g., [63, 55]). Unlike most of the literature in which Green functions are introduced as solutions to partial differential equations (PDE), we systematically derive Green functions from the optimization theory that we have developed in Chapter 9.

[9] This ensures that f is continuous by the Sobolev embedding theorem [11, 28, 3, 102, 132].

13.2.1 Green function of an elliptic PDE

We start with a concrete example and then generalize the derivation to an abstract setting. Consider the elliptic PDE in Example 13.13 in the weak[10] form (13.29) and *for simplicity we also use the classical differential operator ∂ notations (including gradient operator ∇ and the Laplacian operator $\Delta = \nabla \cdot \nabla$) for weak and distributional derivatives, and it will be clear from the context which one we refer to.* We can write (13.29) explicitly as

$$(\nabla u, \nabla v)_{[\mathbb{L}^2(\Omega)]^n} + (u, v)_{\mathbb{L}^2(\Omega)} = (f, v)_{\mathbb{L}^2(\Omega)}, \quad \forall v \in \mathbb{H}_0^1(\Omega). \tag{13.30}$$

As shown in Example 13.13, (13.30) has a unique solution $\tilde{u} \in \mathbb{H}_0^1(\Omega)$.

Suppose now we are interested in evaluating the following quantity of interest—the duality pairing between $\delta \in \mathbb{H}^{-1}(\Omega)$ and $u \in \mathbb{H}_0^1(\Omega)$:

$$J := \delta(u) := \langle \delta, u \rangle_{\mathbb{H}_0^1(\Omega)},$$

which makes sense due to Corollary 13.6. *We present two approaches to evaluate J via the $\mathbb{L}^2$-inner product of f and the Green function of equation (13.25).* The first approach, presented in subsection 13.2.1.1, exploits the Riesz representation Theorem 5.1 since the specific elliptic operator under consideration is associated with $\mathbb{H}^1$-inner product. This approach is not generalizable to general cases, as we shall show in subsection 13.2.2, but provides insights into this specific setting. The second approach in subsection 13.2.1.2 deploys the Lagrangian multiplier Theorem 9.3 which is straightforward to be generalized to abstract settings.

13.2.1.1 Derive the Green function via Riesz representation theorem

Recall from Remark 13.8 that if we can solve the distributional differential equation (13.19), its unique solution is the Riesz representer $g \in \mathbb{H}_0^1(\Omega)$ such that

$$\begin{aligned} J = \left\langle g - \sum_{|\boldsymbol{\alpha}|=1} \mathscr{D}^{\boldsymbol{\alpha}} \left(\mathscr{D}^{\boldsymbol{\alpha}} g \right), u \right\rangle &= \langle g - \Delta g, u \rangle \\ &= (\nabla u, \nabla g)_{[\mathbb{L}^2(\Omega)]^n} + (u, g)_{\mathbb{L}^2(\Omega)} = (f, g)_{\mathbb{L}^2(\Omega)}, \end{aligned}$$

where we have used the weak derivative definition (similar to (13.26)) in the second last equality, and (13.30) in the last equality. As can be seen, if we can solve for the Riesz representer g, we can evaluate J as the inner product of g and f without solving the original PDE (13.30) for u. Now by the Riesz

[10] Again, weak form means considering the PDE using weak derivatives.

representation Theorem 5.1, g satisfies

$$\delta(v) = \langle g - \Delta g, v \rangle, \quad \forall v \in \mathbb{H}_0^1(\Omega),$$

which is equivalent to saying that

$$g - \Delta g = \delta \quad \text{in } \mathbb{H}^{-1}(\Omega). \tag{13.31}$$

A function g that is a solution of (13.31) is known as the Green function for (13.25). Our derivation has shown that the Green function g is a function in $\mathbb{H}^1(\Omega)$. Similar to (13.26), we can write (13.31) in the the following weak form

$$(\nabla g, \nabla \phi)_{[\mathbb{L}^2(\Omega)]^n} + (g, \phi)_{\mathbb{L}^2(\Omega)} = \delta(\phi) = \phi(\mathbf{0}), \quad \forall \phi \in \mathcal{D}(\Omega),$$

where we have used Definition 13.11 for Dirac delta in the last equality.

13.2.1.2 Derive the Green function from optimization

We next consider the following optimization problem

$$\min_{u \in \mathbb{H}_0^1(\Omega)} J, \quad \text{subject to (13.30)}, \tag{13.32}$$

for which the unique solution $\tilde{u}$ of (13.30) is also the optimal solution and thus the optimal quantity of interest is $\tilde{J} = \langle \delta, \tilde{u} \rangle_{\mathbb{H}_0^1(\Omega)}$. Ignoring this obvious fact about the optimal solution, we set out to write down the first-order optimality condition for the optimization problem (13.32) using the Lagrangian Theorem 9.3. In particular, the first-order optimality condition, in fact the adjoint equation (see Remark 9.2), (9.5) reads

$$\delta(h) + (\nabla h, \nabla v)_{[\mathbb{L}^2(\Omega)]^n} + (h, v)_{\mathbb{L}^2(\Omega)} = 0, \quad \forall h \in \mathbb{H}_0^1(\Omega), \tag{13.33}$$

for some $v \in \mathbb{H}_0^1(\Omega)$. By the density of $\mathcal{D}(\Omega)$ in $\mathbb{H}_0^1(\Omega)$ (see Definition 13.16), we have

$$\delta(\phi) + (\nabla h, \nabla \phi)_{[\mathbb{L}^2(\Omega)]^n} + (h, \phi)_{\mathbb{L}^2(\Omega)} = 0, \quad \forall \phi \in \mathcal{D}(\Omega),$$

which, by using Dirac delta definition (13.11) for the first term, distributional derivative (13.12) for the second term, and the regular distribution (13.7) for the last term, can be written equivalently as

$$-\Delta v + v = -\delta, \tag{13.34}$$

in the distributional sense.

A few observations are in order for the adjoint equation (13.34). First, the adjoint solution v in this case is known as the Green function for the PDE (13.25) and we have derived it from the Lagrangian multiplier Theorem 9.3. Second, even when the original PDE has a classical solution u, the Green function v is a distribution in general as the Green equation (13.34) is only valid in the distributional sense. Specifically, the Lagrangian multiplier Theorem 9.3 tells us that $v \in \mathbb{H}_0^1(\Omega)$ and this makes perfect sense for the Green equation (13.34), as then the left hand side $-\Delta v + v$ is a distribution in $\mathbb{H}^{-1}(\Omega)$ (see Example 13.10) and the right hand side δ is also a distribution in $\mathbb{H}^{-1}(\Omega)$ by Corollary 13.6. In other words, (13.34) is an equality in $\mathbb{H}^{-1}(\Omega)$. The result is identical—up to a sign difference which is immaterial and can be corrected by defining the constraint with the negative sign—and consistent with the derivation of the Green function from the Riesz representation Theorem 5.1 in subsection 13.2.1.1.

The beauty of the Green function is that, once determined, it can be used as the inverse operator of the original differential operator to compute the solutions of the corresponding PDE for any new right-hand side f in (13.30). Indeed, from the definition of J and the Green function (13.34) we have

$$J = \delta(u) = -(\nabla u, \nabla v)_{[\mathbb{L}^2(\Omega)]^n} - (u, v)_{\mathbb{L}^2(\Omega)} = -(f, v)_{\mathbb{L}^2(\Omega)} \tag{13.35}$$

where we have used (13.33) in the second equality with $h = u$, and (13.30) in the last equality. As a result, evaluating J for an arbitrary f amounts to computing the $\mathbb{L}^2$-inner product of the Green function v and f.

13.2.2 Green functions of general linear operators

Let us now consider an abstract linear operator $\mathscr{A} : \mathsf{D}(\mathscr{A}) \underset{\text{dense}}{\subset} \mathbb{X} \to \mathbb{Y}^*$, with X and Y being Hilbert spaces, and the following equation

$$\mathscr{A} u = f, \tag{13.36}$$

which is assumed to have a unique solution $\tilde{u}$ in $\mathbb{X}$ for every $f \in \mathbb{Y}^*$. Let us denote by $\mathscr{A}^{-1} : \mathbb{Y} \to \mathbb{X}^*$ the inverse of $\mathscr{A}$. We also define the adjoint $\mathscr{A}^* : \mathbb{Y} \to \mathbb{X}^*$ as

$$\langle \mathscr{A} v, w \rangle_{\mathbb{Y}} = \langle v, \mathscr{A}^* w \rangle_{\mathbb{X}}, \quad \forall v \in \mathsf{D}(\mathscr{A}) \text{ and } w \in \mathsf{D}(\mathscr{A}^*), \tag{13.37}$$

where

$$\mathsf{D}(\mathscr{A}^*) := \{w : \text{ the map } v \mapsto \langle \mathscr{A} v, w \rangle_{\mathbb{Y}} \text{ is continuous on } \mathbb{X}\}. \tag{13.38}$$

Similar to section 12.1, we can justify the above definition of adjoint (see Problem 13.4) in particular one can show that $\left(\mathscr{A}^{-1}\right)^* = (\mathscr{A}^*)^{-1}$. We would

like to point out that the setting in (13.27), and hence subsection 13.2.1, is a special case of this section when we take $\mathbb{X} := \mathbb{H}_0^1(\Omega)$ and $\mathbb{Y}^* \equiv \mathbb{Y} := \mathbb{L}^2(\Omega)$. Suppose that the Dirac delta δ is well-defined[11] on $\mathbb{X}$ and we are interested in evaluating the following duality pairing

$$J := \langle \delta, u \rangle_{\mathbb{X}}.$$

By assumption, δ is a linear and continuous functional on $\mathbb{X}$, and by the Riesz representation Theorem 5.1 there is a unique $g \in \mathbb{X}$ such that

$$\langle \delta, v \rangle_{\mathbb{X}} = (v, g)_{\mathbb{X}}, \quad \forall v \in \mathbb{X},$$

which is the equation to solve for g. Unlike a specific setting in subsection 13.2.1.1 in which we can write an explicit equation for g in (13.31), all we can say here is that the unique g can be solved, at least in principle, once we are given the $\mathbb{X}$-inner product. We then have

$$J = \langle \delta, u \rangle_{\mathbb{X}} = (u, g)_{\mathbb{X}} = \left(\mathscr{A}^{-1} f, g\right)_{\mathbb{X}} = \left\langle f, \left(\mathscr{A}^*\right)^{-1} g \right\rangle_{\mathbb{Y}},$$

where we have used duality pairings to define the adjoint of $\mathscr{A}^{-1}$ in the last equality. We have shown that we can evaluate J for any f by determining two quantities: i) the Riesz representation g of δ in $\mathbb{X}$ and ii) the inverse of the adjoint operator $\mathscr{A}^*$. Since the operator $\mathscr{A}$ is not related to the $\mathbb{X}$-inner product in general, g is not related to the Green function associated with $\mathscr{A}$ in general. In other words, the approach relies on the Riesz representation theorem in subsection 13.2.1.1 fails to lead to the Green function for general linear operators.

We begin the Lagrangian approach by considering the following trivial optimization problem

$$\min_{u \in \mathbb{X}} J(u), \quad \text{subject to (13.36)}, \tag{13.39}$$

for which the unique solution $\tilde{u}$ of (13.36) is also the optimal solution and thus the optimal quantity of interest is $\tilde{J} = \langle \delta, \tilde{u} \rangle_{\mathbb{X}}$. Ignoring this obvious fact about the optimal solution, we write down the first-order optimality condition for the optimization problem (13.39) using the Lagrangian Theorem 9.3. The adjoint equation (9.5) (see Remark 9.2) reads

$$\delta(h) + \langle \mathscr{A} h, v \rangle_{\mathbb{Y}} = 0, \quad \forall h \in \mathbb{X}, \tag{13.40}$$

which, after using the definition of adjoint in (13.37) and the assumption that δ is a linear and continuous on $\mathbb{X}$, becomes

$$\langle h, \delta \rangle_{\mathbb{X}} + \langle h, \mathscr{A}^* v \rangle_{\mathbb{X}} = 0, \quad \forall h \in \mathbb{X},$$

[11] From Corollary 13.6, $\mathbb{X}$ should be some Hilbert subspace of $\mathbb{H}_0^1(\Omega)$.

which is equivalent to

$$\mathscr{A}^* v = \delta \tag{13.41}$$

as an equation in $\mathbb{X}^*$. *Such a v is know as the Green function for the equation (13.36).*

Our systematic and constructive derivation of the Green function induces several important properties of the Green function. First, the existence of the Green function v is guaranteed by the Lagrangian Theorem 9.3. In particular, the adjoint/Green equation (13.41) guarantees to have at least one Green function v as its solution. Second, we have, again, shown that the Green function is nothing more than the adjoint solution resulting from the first-order optimality condition of the optimization problem (13.39), namely, the adjoint of (13.36). Third, our approach also shows that the Green function is a member of $\mathsf{D}(\mathscr{A}^*) \subset \mathbb{Y}$. Fourth, we can now express the objective function J in terms of the Green function v and the forcing term f as follow

$$J = \delta(u) = -\langle \mathscr{A} u, v \rangle_{\mathbb{Y}} = -\langle f, v \rangle_{\mathbb{Y}},$$

where we have used (13.40) in the second equality and (13.36) in the last equality. When $\mathbb{Y} = \mathbb{L}^2(\Omega)$, we recover the result in (13.35).

A generalization to any linear functional $\ell \in \mathbb{X}^*$, and thus not requiring that $\mathbb{X}$ be a Hilbert subspace of $\mathbb{H}^1(\Omega)$, is straightforward and is left as an exercise (see Problem 13.6).

Problems

Problem 13.1. Show that the Dirac delta function defined via (13.8) does not reside in any $\mathbb{L}^p$ function spaces for $p \geq 1$.

Problem 13.2. Prove Proposition 13.1.

Problem 13.3. Let m and p be two arbitrary integers. Show that $\mathbb{H}^m(\Omega) \subset \mathbb{H}^p(\Omega)$ when $m > p \geq 0$ or $p < m \leq 0$.
Hint. For $m > p \geq 0$, the conclusion is oblivious by Definition 13.15 of the Sobolev spaces. Now consider $p < m \leq 0$ and take $u \in \mathbb{H}^p(\Omega)$ and $\varphi \in \mathbb{H}^{-p}$. By Definition 13.15, we have $\varphi \in \mathbb{H}^{-m}$, and thus

$$|\langle u, \varphi \rangle_{\mathbb{H}^{-m}}| \leq \|u\|_{\mathbb{H}^m} \|\varphi\|_{\mathbb{H}^{-m}} \leq \|u\|_{\mathbb{H}^m} \|\varphi\|_{\mathbb{H}^{-p}},$$

and thus $u \in \mathbb{H}^p(\Omega)$.

Problem 13.4. Following section 12.1, show the existence and uniqueness of the adjoint operator defined in (13.37) with the domain defined in (13.38). In addition, show that $\left(\mathscr{A}^{-1}\right)^* = \left(\mathscr{A}^*\right)^{-1}$.

Problem 13.5. Let $\Omega \subset \mathbb{R}^n$ be open and bounded. Consider the following elliptic PDE:

$$\begin{aligned} -\Delta u &= f, \quad \text{in } \Omega, \\ u &= 0, \quad \text{in } \partial\Omega, \end{aligned}$$

where $\Delta(\cdot) := \sum_{i=1}^{n} \partial_{\boldsymbol{x}_i}(\partial_{\boldsymbol{x}_i}(\cdot))$ is the Laplacian. Following Example 13.13, derive the weak formulation for the above PDE. In Example 15.4, we will show that the weak setting has a unique solution. Derive the Green function using both the Riesz representation theorem and optimization method. Which one gives the standard Green function? Why doesn't the other?

Hint. We do not get the same result because the Riesz representation approach does not actually give the Green function associated with the PDE. The PDE is now no longer associated with the Hs^1-inner product. Thus, though we can still get through the result with the Riesz representation theorem, it does not yield the standard Green function.

Problem 13.6. Consider the setting in subsection 13.2.2, but not instead of δ we consider a general linear function $\ell \in \mathbb{X}^*$, and thus not requiring that $\mathbb{X}$ be a Hilbert subspace of $\mathbb{H}^1(\Omega)$. Find the equation for a generalized Green function v such that

$$J := \langle \ell, u \rangle_{\mathbb{X}} = -\langle f, v \rangle_{\mathbb{Y}},$$

where u is a solution of (13.36).

Chapter 14
Understanding ill-posed problems using the singular value decomposition

Abstract

In this section, we consider continuous linear operator defined on the whole space and we shall extend the spectral decomposition in Corollary 7.1 and SVD decomposition in Theorem 8.1 to compact (linear) operators in infinite dimensions. This allows us to show that inverting a compact operator is an ill-posed problem. We then explain rigorously how the standard Tikhonov regularization could overcome the ill-posedness. We begin by recalling the definition of compact linear operators and some of its consequences.

14.1 Preliminary

Definition 14.1 (Compact operator). Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be linear. We say that $\mathscr{A}$ is a compact operator if for every bounded sequence $\{u_i\}_{i=1}^{\infty} \subset \mathbb{X}$, the sequence $\{\mathscr{A}u_i\}_{i=1}^{\infty} \subset \mathbb{Y}$ has a convergent subsequence.

A direct consequence of Definition 14.1 is that any compact operator is a linear and continuous map.

Corollary 14.1. *If $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ is compact, then $\mathscr{A} \in \mathcal{B}(\mathbb{X}, \mathbb{Y})$.*

14.2 A version of the Hilbert-Schmidt theorem

Self-adjoint compact operators in Hilbert spaces possesses many desirable properties among which are countable real spectrum, finite dimensional eigenspaces for non-zero eigenvalues, and the convergence to zero of eigenvalues when the number of them are infinite (see, e.g., [11, 28, 112, 126]). One of the important consequences is the Hilbert-Schmidt theorem (see, e.g., [11, 28, 112, 126, 131]), which is a generalization of Corollary 7.1.

Theorem 14.1 (Hilbert-Schmidt theorem for self-adjoint compact operators: first version). *Let $\mathscr{B} : \mathbb{X} \to \mathbb{X}$ be a self-adjoint compact operator. Then there exists an orthonormal set of eigen-functions φ_i corresponding to non-zero eigenvalues λ_i of $\mathscr{B}$ such that for any $x \in \mathbb{X}$ we have a unique expansion of the form*

$$x = \sum_i (\varphi_i, x)_{\mathbb{X}} \varphi_i + \mathscr{P}x, \tag{14.1}$$

where $\mathscr{P}$ is an orthogonal projection from $\mathbb{X}$ to the nullspace $\mathsf{N}(\mathscr{B})$.

Furthermore, we have

$$\mathscr{B}x = \sum_i \lambda_i (\varphi_i, x)_{\mathbb{X}} \varphi_i,$$

that is, the set of all eigenfunctions $\{\varphi_i\}$ forms a basis for $\mathsf{R}(\mathscr{B})$, and the convergence for the series on the right hand side is in the $\mathbb{X}$-topology.

Remark 14.1. Note that the spectral expansion of x in (14.1) can be considered as a generalization of the Fourier series in (16.1). Even when $\mathbb{X}$ is not countable, $\mathscr{P}x$ can be expanded in a countable orthonormal set that is orthogonal to all eigenfunctions (see [108, Lemma 5.17.17]).

Remark 14.2. The fact that $\mathscr{B}$ is compact when $\mathbb{X}$ is a finite dimensional space implies that Theorem 14.1 is a generalization of Corollary 7.1. Indeed, let $\dim(\mathbb{X}) = n$. In this case, the eigenspace corresponding to the zero eigenvalue is spanned by finite number of (say d) orthonormal eigenfunctions $\{\varphi_j^0\}_{j=1}^d$ and thus $\mathscr{P}x = \sum_{j=1}^{d} (\varphi_j^0, x)_{\mathbb{X}} \varphi_j^0$, which can be absorbed into the first sum on the right side of (14.1) so that we can write

$$x = \sum_{i=1}^{n} (\varphi_i, x)_{\mathbb{X}} \varphi_i,$$

after renaming the eigenfunctions corresponding to zero eigenvalues. This is exactly Corollary 7.1.

- Since it is now a book, we should prove the Hilbert-Schmidt theorem? Look back at the book by Showalter again for the proof.
- Should separate the rest of the chapter in a new chapter "Adjoint in infinite dimensional SVD and application to ill-posed problems.
- perhaps add another chapter on SVD for non-symmetric kernel and the error bound for neural networks with kernel approach as in the paper "Approximation by non-symmetric networks for cross-domain learning" by H. N. Mhaskar? I think the unifying paper "Green's Functions: Taking Another Look at Kernel Approximation, Radial Basis Functions and

Splines" by Gregory E. Fasshauer is nice. We should expand this chapter based on this paper?

14.3 SVD of compact operators in Hilbert spaces

We now follow the exposition in Chapter 8 to construct the SVD for compact operators. Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be a compact operator, then it can be shown that that $\mathscr{B} := \mathscr{A}^*\mathscr{A} : \mathbb{X} \to \mathbb{X}$ is a self-adjoint compact operator [112, 11]. The Hilbert-Schmidt Theorem 14.1 says that there exists orthonormal eigenfunctions φ_i corresponding to nonzero eigenvalues λ_i of $\mathscr{B}$ such that

$$\mathscr{B}\varphi_i = \mathscr{A}^*\mathscr{A}\varphi_i = \lambda_i\varphi_i,$$

which implies

$$(\mathscr{A}^*\mathscr{A}\varphi_i, \varphi_i)_{\mathbb{X}} = \lambda_i (\varphi_i, \varphi_i)_{\mathbb{X}},$$

which, by definition of $\mathscr{A}^*$, in turn can be written as

$$\|\mathscr{A}\varphi_i\|_{\mathbb{X}}^2 = \lambda_i \|\varphi_i\|_{\mathbb{X}}^2,$$

which shows that $\lambda_i \geq 0$. Let us define the *singular value* σ_i of $\mathscr{A}$ as

$$\sigma_i := \sqrt{\lambda_i}. \tag{14.2}$$

We are now in the position to study the singular value decomposition for compact operators (see, e.g.,[37]) that is a direct extension of Theorem 8.1.

Theorem 14.2 (Singular value decomposition for compact operators). *Let $\{\sigma_i\}$ be the sequence of non-zero singular values (defined in (14.2)) of a compact operator $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ and be ordered as*

$$\sigma_1 \geq \sigma_2 \geq \ldots,$$

then there exist two orthonormal sequences $\{\varphi_i\}$ and $\{\phi_i\}$ such that

1. *$\mathscr{A}\varphi_i = \sigma_i\phi_i$ and $\mathscr{A}^*\phi_i = \sigma_i\varphi_i$.*
2. *$\forall x \in \mathbb{X}$, we have*
$$x = \sum_i (x, \varphi_i)_{\mathbb{X}} \varphi_i + \mathscr{P}\varphi,$$
$\mathscr{P} : \mathbb{X} \to \mathsf{N}(\mathscr{A})$ is an orthogonal projection.
3. *There holds*
$$\mathscr{A}x = \sum_i \sigma_i (x, \varphi_i)_{\mathbb{X}} \phi_i.$$

We call $\{\sigma_i, \varphi_i, \phi_i\}$, $i = 1, 2, \ldots$, **the singular system** *of $\mathscr{A}$.*

Proof. The proof of this theorem is similar to its finite dimensional counterpart Theorem 8.1. The key that we exploit is the Hilbert-Schmidt theorem Theorem 14.1.

1. By Theorem 14.1 we have

$$\mathscr{A}^*\mathscr{A}\varphi_i = \sigma_i^2\varphi_i,$$

where $\{\varphi_i\}$ is an orthonormal set in $\mathbb{X}$. Let us define

$$\sigma_i\phi_i := \mathscr{A}\varphi_i,$$

then

$$(\phi_i,\phi_j)_{\mathbb{Y}} = \frac{1}{\sigma_i\sigma_j}(\mathscr{A}\varphi_i,\mathscr{A}\varphi_j)_{\mathbb{Y}} = \frac{1}{\sigma_i\sigma_j}(\mathscr{A}^*\mathscr{A}\varphi_i,\varphi_j)_{\mathbb{X}} = \frac{\sigma_i}{\sigma_j}(\varphi_i,\varphi_j)_{\mathbb{X}} = \delta_{ij}.$$

That is, $\{\phi_i\}$ is an orthonormal set in $\mathbb{Y}$. By definition we have

$$\mathscr{A}^*\phi_i = \frac{1}{\sigma_i}\mathscr{A}^*\mathscr{A}\varphi_i = \sigma_i\varphi_i.$$

2. Again, by Hilbert-Schmidt Theorem 14.1 we have

$$\forall x\in\mathbb{X}: \quad x = \sum (x,\varphi_i)_{\mathbb{X}}\varphi_i + \mathscr{P}\varphi,$$

where $\mathscr{P}$ is an orthonormal projection from $\mathbb{X}$ to $\mathsf{N}(\mathscr{A}^*\mathscr{A})$. The second assertion is now clear owing to the fact that $\mathsf{N}(\mathscr{A}) = \mathsf{N}(\mathscr{A}^*\mathscr{A})$.
3. We start with the partial sum

$$s_N := \sum_{i=1}^{N}(x,\varphi_i)_{\mathbb{X}}\varphi_i,$$

and thus

$$\mathscr{A}s_N = \sum_{i=1}^{N}\sigma_i(x,\varphi_i)_{\mathbb{X}}\phi_i.$$

Now passing to the limit we obtain

$$\lim_{N\to\infty}\mathscr{A}s_N = \mathscr{A}(x-\mathscr{P}x) = \mathscr{A}x.$$

Consequently,

$$\mathscr{A}x = \sum_i \mu_i(x,\varphi_i)_{\mathbb{X}}\phi_i.$$

The Theorem 14.2 provides a trivial proof for the following result.

Corollary 14.2. *Let $\mathscr{A}:\mathbb{X}\to\mathbb{Y}$ be a compact operator and its singular triplets are given in Theorem 14.2. The following hold.*

1. *The SVD of* $\mathscr{A}^* : \mathbb{Y} \to \mathbb{X}$ *is given as*

$$\mathscr{A}^* y = \sum_i \sigma_i \, (y, \phi_i)_{\mathbb{Y}} \, \varphi_i,$$

for any $y \in \mathbb{Y}$.

2. φ_1 *and* ϕ_1 *are, respectively, a solution of the following problem*

$$\|\mathscr{A}\| := \sup_{x \in \mathbb{X}} \frac{\|\mathscr{A} x\|_{\mathbb{X}}}{\|x\|_{\mathbb{X}}}, \qquad \textit{and} \;\; \|\mathscr{A}^*\| := \sup_{y \in \mathbb{Y}} \frac{\|\mathscr{A}^* y\|_{\mathbb{Y}}}{\|y\|_{\mathbb{Y}}},$$

and $\|\mathscr{A}\| = \|\mathscr{A}^*\| = \sigma_1$.

Proof. The proof for the first assertion is easy and left as an exercise in Problem 14.1. For the second assertion, the proof is similar to the proof of Corollary 8.1. Recall from Theorem 14.2 that any $x \in \mathbb{X}$ can be expressed as

$$x = \sum_i (x, \varphi_i)_{\mathbb{X}} \, \varphi_i + \mathscr{P} \varphi,$$

and by Bessel inequality we have

$$\|\mathscr{A} x\|_{\mathbb{Y}}^2 = \sum_i \sigma_i^2 \, |(x, \varphi_i)_{\mathbb{X}}|^2 \leq \sigma_1^2 \sum_i |(x, \varphi_i)_{\mathbb{X}}|^2 \leq \sigma_1^2 \, \|x\|_{\mathbb{X}}^2 < \infty,$$

and the equality happens when $x = \varphi_1$. Furthermore, in that case

$$\sup_{x \in X} \frac{\|\mathscr{A} x\|_{\mathbb{Y}}}{\|x\|_{\mathbb{X}}} = \sigma_1.$$

The next result [37], due to Picard, tells us the conditions under which inverting a compact operator is well-defined. *As we will see, the adjoint plays a key role.*

Theorem 14.3 (Picard). *Suppose* $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ *is a compact operator. The equation*

$$\mathscr{A} x = y$$

is solvable iff

i) $y \in \mathsf{N}\,(\mathscr{A}^*)^{\perp}$, *and*
ii) $\sum \dfrac{1}{\sigma_i^2} \, |(y, \phi_i)_{\mathbb{Y}}|^2 < \infty,$

where $\{\sigma_i, \varphi_i, \phi_i\}$ *is the singular system of* $\mathscr{A}$. *In this case the solution is given by*

$$x = \sum_i \frac{1}{\sigma_i} \, (y, \phi_i)_{\mathbb{Y}} \, \varphi_i.$$

Proof. The SVD Theorem 14.2 provides a simple proof for this theorem.

$\Rightarrow$ Solvability implies that y belongs to the range of $\mathscr{A}$, i.e. $y \in \mathsf{R}(\mathscr{A})$. From the closed range Theorem 5.2 [1] we know that $\mathsf{R}(\mathscr{A}) \subset \overline{\mathsf{R}(\mathscr{A})} = \mathsf{N}(\mathscr{A}^*)^\perp$, and hence *i)* holds. On the other hand, by Theorem 14.1 we can express a solution x as

$$x = \sum_i (x, \varphi_i)_{\mathbb{X}} \varphi_i + \mathscr{P} x,$$

which, together with the Parseval identity, implies

$$\|x\|_{\mathbb{X}}^2 = \sum_i |(x, \varphi_i)_{\mathbb{X}}|^2 + \|\mathscr{P} x\|_{\mathbb{X}}^2, \tag{14.3}$$

which in turns implies

$$\sum_i |(x, \varphi_i)_{\mathbb{X}}|^2 \leq \|x\|_{\mathbb{X}}^2 < \infty.$$

Since

$$(x, \varphi_i)_{\mathbb{X}} = \frac{1}{\sigma_i} (x, \mathscr{A}^* \phi_i)_{\mathbb{X}} = \frac{1}{\sigma_i} (\mathscr{A} x, \phi_i)_{\mathbb{Y}} = \frac{1}{\sigma_i} (y, \phi_i)_{\mathbb{Y}},$$

the assertion *ii)* holds.

$\Leftarrow$ Since $y \in \mathsf{N}(\mathscr{A}^*)^\perp$, Hilbert-Schmidt Theorem 14.1 gives

$$y = \sum_i (y, \phi_i)_{\mathbb{Y}} \phi_i.$$

Now, from *ii)* the following definition

$$x := \sum_i \frac{1}{\sigma_i} (y, \phi_i)_{\mathbb{Y}} \varphi_i$$

is meaningful. Together with Corollary 14.1, we have

$$\mathscr{A} x = \sum_i \frac{1}{\sigma_i} (y, \phi_i)_{\mathbb{Y}} \mathscr{A} \varphi_i = \sum_i (y, \phi_i)_{\mathbb{Y}} \phi_i = y,$$

where we have used in the last equality the Hilbert-Schmidt Theorem 14.1 for $\mathscr{B} = \mathscr{A}\mathscr{A}^*$, the fact that ϕ_i are eigenfunctions of $\mathscr{B}$, and $y \in \mathsf{N}(\mathscr{A}^*)^\perp$. This concludes the proof.

[1] Note that $\mathsf{R}(\mathscr{A})$ cannot be closed since $\mathscr{A}$ is compact. Assume, on the contrary, it is, then by the bounded inverse theorem [127] we know that $\mathscr{A}^{-1}$ is continuous and hence $\mathscr{I} = \mathscr{A}^{-1}\mathscr{A}$ is also a compact operator. But, this is a contradiction since identity operator in infinite dimensional space cannot be a compact operator [112, 11]. If $R(\mathscr{A})$ were closed, then *i)* would be both necessary and sufficient. Since this is not true for compact operators, we have to replace the closedness by the smooth property *ii)* of the right hand side.

We now discuss the important consequence of the Picard Theorem 14.3, that is, *inverting a compact operator is an ill-posed problem*. We observe that

$$y = \mathscr{A}x = \sum_i \sigma_i \left(x, \phi_i\right)_{\mathbb{X}} \phi_i,$$

where we have used the second assertion of Theorem 14.2. Since $\mathscr{A}$ is compact, and hence $\sigma_i \to 0$ as $i \to \infty$, $\mathscr{A}$ smoothes out the contribution from the "high frequency" mode: i.e. φ_i for large i. In other words, the output y is insensitive to high frequency modes φ_i when $i \to \infty$.

Conversely, let us perturb the right hand side y as

$$\tilde{y} = y + \delta\phi_N,$$

where $\delta \in \mathbb{R}$ and $N \in \mathbb{N}$, then the corresponding solution reads

$$\tilde{x} = \sum_i \frac{1}{\mu_i} \left(\tilde{y}, \phi_i\right)_{\mathbb{Y}} \varphi_i = x + \frac{\delta}{\mu_N}\varphi_N.$$

Thus,

$$\frac{\|\tilde{x} - x\|_{\mathbb{X}}}{\|\tilde{y} - y\|_{\mathbb{Y}}} = \frac{1}{\mu_N} \to \infty, \text{ as } N \to \infty,$$

which is exactly the subtle instability problem of inverting a compact operator, namely, small changes in the input can lead to very large change in the solution. Most of linear inverse problems (such as deconvolution) fall into this category, and practical nonlinear inverse problems do too (see, e.g., [34, 33, 32] and the references therein).

Definition 14.2 (Well-posedness). In Hadamard's sense [65], the problem $\mathscr{A}x = g$ is well-posed if

1. $\mathscr{A}$ is surjective (*there exists a solution*: **existence**),
2. $\mathscr{A}$ is injective (*there is at most one solution*: **uniqueness**), and
3. $\mathscr{A}^{-1}$ is continuous (*the solution depends continuously on the data*: **stability**).

Example 14.1 (Inverse of the fundamental theorem of calculus). Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ and consider the fundamental theorem of calculus in the following form

$$y\left(t\right) = \mathscr{A}x := \int_0^t x\left(s\right)\, ds, \quad 0 \le t \le 1,$$

and the inverse problem is to find x given its anti-derivative y. We are going to show that, depending on $\mathbb{X}, \mathbb{Y}$, this inverse problem can be ill-posed or well-posed.

- First let us consider $\mathbb{X} = \mathcal{C}\left([0,1]\right), \mathbb{Y} = \mathcal{C}\left([0,1]\right)$. Let $\mathscr{A}x = y$ and consider

$$\tilde{y} := y - \frac{\alpha}{N} + \frac{\delta}{N}\cos(Nt).$$

Then, by the fundamental theorem of calculus, the corresponding solution is given by

$$\tilde{x} = x + \delta \sin(Nt).$$

Clearly

$$\|\tilde{y} - y\|_{\mathcal{C}([0,1])} := \sup_{t\in[0,1]} |\tilde{y}(t) - y(t)| \to 0, \quad \text{as } N \to \infty,$$

but

$$\|\tilde{x} - x\|_{\mathcal{C}([0,1])} = \alpha \quad \forall N.$$

We conclude that $\mathscr{A}$ does not distinguish x and $\tilde{x}$, and as the result the inverse problem does not have a unique solution. In fact, $\mathscr{A}$ is a compact operator[2] and, as we have discussed above, it "smoonthes" out the difference in x and $\tilde{x}$ so that the observation y is the same. Intuitively, a compact operator "squeezes" its domain into "smaller" range: for the above example $\mathscr{A}$, as an integral operator, maps $\mathcal{C}([0,1])$ into $\mathcal{C}^1([0,1]) \subset \mathcal{C}([0,1])$. Since the inverse of a compact operator is unbounded, inverting the fundamental theorem of calculus is unstable by the Picard Theorem 14.3. The setting $\mathbb{X} = \mathcal{C}([0,1]), \mathbb{Y} = \mathcal{C}([0,1])$ thus leads to an ill-posed problem.

- Now let us consider $\mathbb{X} = \mathcal{C}([0,1]), \mathbb{Y} = \mathcal{C}^1([0,1])$. In this case we have

$$\|\tilde{y} - y\|_{\mathcal{C}^1([0,1])} := \|\tilde{y} - y\|_{\mathcal{C}([0,1])} + \|\tilde{y}' - y'\|_{\mathcal{C}([0,1])} = \alpha, \quad \forall N,$$

and since

$$\|\tilde{x} - x\|_{\mathcal{C}([0,1])} = \alpha \quad \forall N.$$

we conclude that a small change in y leads to a small (in fact the same) change in x. The inverse problem is thus stable. The uniqueness is also trivial due to the fact that $\frac{dy}{dt} = x$. The surjectivity is also clear. Consequently, the inverse problem is well-posed in the Hadamard's sense.[3]

Remark 14.3. In practice, we do not solve $\mathscr{A}x = y$ directly on the infinite dimensional setting but via some discretization approach to obtain a finite dimensional problem to solve (on computer). This does not go around the ill-posedness issue. Indeed, in this case, the compactess of $\mathscr{A}$ is manifest in the ill-conditioning of its discrete counterpart whose smallest singular value could be very small. Inverting the discrete system is thus an ill-conditioned problem—a discrete way of saying ill-posedness.

[2] By the Ascoli-Arzela theorem [112, 11].

[3] Note that this is an instance of the Tikhonov theorem [46] since $\mathcal{C}^1([0,1])$ is compactly embedded in $\mathcal{C}([0,1])$.

We have seen that there could be multiple (or there is no) solutions to the linear problem of interest $\mathscr{A}x = y$. The main reason is that the nullspace of $\mathscr{A}$ is non-trivial or y is not in the range of $\mathscr{A}$. The question is if we can find a "useful solution" in this case? One way to address this question is to look for the solution that minimizes the residual, such as the least squares problem in Corollary 11.1:

$$\min_x \frac{1}{2} \|\mathscr{A}x - y\|_{\mathbb{Y}}^2 . \tag{14.4}$$

However, when $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ is compact, the ill-posedness nature of our inverse problem does not go away as the normal equation (11.1) is still ill-posed due to the fact that $\mathscr{A}^*\mathscr{A}$ is compact. In other words, we still have problem with the uniqueness if $\mathsf{N}(\mathscr{A})$ is not trivial, and the (bigger) problem with instability due to inverting the compact operator $\mathscr{A}^*\mathscr{A}$. However, the optimization idea paves the way for using optimization technique to overcome the ill-posedness problem, as we now discuss. Note that the objective function in (14.4) is quadratic in x, and hence a "parabola". Clearly, if it is a well-behaved parabola, then the minimizer is unique. This immediately suggests that we should add a quadratic term to the objective function to improve its behavior, and hence removing the uniqueness issue: as will be shown, this also addresses the stability. This is essentially the idea behind the *Tikhonov regularization* [138, 137], which proposes to solve the following nearby problem

$$\min_{x \in \mathbb{X}} \frac{1}{2} \|\mathscr{A}x - y\|_{\mathbb{Y}}^2 + \frac{\kappa}{2} \|x - x_0\|_{\mathbb{X}}^2 , \tag{14.5}$$

where x_0 is some "prior" reference function and κ is known as the *regularization parameter*. To show that the *regularized optimization* problem (14.5) is well-posed, we need the projection theorem Theorem 11.1 and the following key result from the Riesz-Fredholm theory [36].

Lemma 14.1. *Let $\mathscr{A}$ be a compact operator from $\mathbb{X}$ to $\mathbb{X}$. If $(I + \mathscr{A})$ is injective, then $(I + \mathscr{A})$ is continuously invertible.*

Theorem 14.4. *For any $\kappa > 0$, the regularized optimization problem (14.5) is well-posed.*

Proof. Without loss of generality, assume $\kappa = 1$. We begin by rewrite the optimization (14.5) into the following equivalent form

$$\min_z \frac{1}{2} \|\mathscr{B}z - w\|_{\mathbb{Y}\times\mathbb{X}}^2 , \tag{14.6}$$

where we have defined $\mathscr{B} : \mathbb{X} \ni z \mapsto [\mathscr{A}, \mathscr{I}]\, z := [\mathscr{A}z, z] \in \mathbb{Y} \times \mathbb{X}$, and $w := [w_1, w_2] := [y, z_0]$. The inner product of $z, w \in \mathbb{Y}\times\mathbb{X}$ is defined as $(z, w)_{\mathbb{Y}\times\mathbb{X}} := (z_1, w_1)_{\mathbb{Y}} + (z_2, w_2)_{\mathbb{X}}$, and the induced norm for any $z = [z_1, z_2] \in \mathbb{Y} \times \mathbb{X}$ is given by $\|z\|_{\mathbb{Y}\times\mathbb{X}}^2 := \|z_1\|_{\mathbb{Y}}^2 + \|z_2\|_{\mathbb{X}}^2$. From the definition of the inner product in $\mathbb{Y} \times \mathbb{X}$, the definition of adjoint, and the fact that the identity operator

$\mathscr{I}$ is self-adjoint, we have $\mathscr{B}^* z = \mathscr{A}^* z_1 + z_2$. Next, from Corollary 11.1 we know that the minimizer satisfies

$$\mathscr{B}^* \mathscr{B} z = \mathscr{B}^* w,$$

which is equivalent to

$$\left(\mathscr{A}^* \mathscr{A} + I\right) z = \mathscr{A}^* y + z_0.$$

Since $\left(\mathscr{A}^* \mathscr{A} + I\right)$ is injective[4], Lemma 14.1 shows that it is continuously invertible, i.e. $\left\| \left(\mathscr{A}^* \mathscr{A} + I\right)^{-1} \right\| < \infty$. Hence,

$$\|z\|_{\mathbb{X}} = \left\| \left(\mathscr{A}^* \mathscr{A} + I\right)^{-1} \left(\mathscr{A}^* y + z_0\right) \right\| \leq \left\| \left(\mathscr{A}^* \mathscr{A} + I\right)^{-1} \right\| \left(\|\mathscr{A}^*\| \, \|y\|_{\mathbb{X}} + \beta \, \|z_0\|_{\mathbb{X}} \right),$$

that is, the solution x of the Tikhonov regularization (14.5) is not only unique but also depends continuously on the data y, and this concludes the proof.

Problems

Problem 14.1. Prove the first assertion of Corollary 14.2.

[4] From $\left(\mathscr{A}^* \mathscr{A} + I\right) z = 0$ we have $0 = \left(x, \left(\mathscr{A}^* \mathscr{A} + I\right) z\right)_{\mathbb{X}} = \|\mathscr{A} x\|_{\mathbb{Y}}^2 + \|x\|_{\mathbb{X}}^2$ and thus $x = \theta$.

Chapter 15
Wellposedness of linear operator equation via adjoint

Abstract

In this section we are interested in the well-posedness (in the sense of Hadamard in Definition 14.2) of operator equation $\mathscr{A}x = y$, where $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ is linear and continuous. Our exposition follows [51] closely. We begin with a key result [11, 7, 51].

Lemma 15.1. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be linear and continuous. Then*

$$\mathscr{A} \text{ is bounded below } \Leftrightarrow \exists \alpha > 0 : \|\mathscr{A}u\|_{\mathbb{Y}} \geq \alpha \|u\|_{\mathbb{X}} \Leftrightarrow \begin{cases} \mathscr{A} \text{ is injective,} \\ \mathsf{R}(\mathscr{A}) \text{ is closed.} \end{cases}$$

Proof. We provide a proof in section 15.1.

The following result highlights the role of the adjoint operator $\mathscr{A}^*$ on the injectivity and the closedness of $\mathsf{R}(\mathscr{A})$, and hence the boundedness below of $\mathscr{A}$.

Theorem 15.1. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be linear and continuous. The following are equivalent:*

1) $\mathscr{A}^* : \mathbb{Y} \to \mathbb{X}$ is surjective.
2) $\mathscr{A}$ is injective and $\mathsf{R}(\mathscr{A})$ is closed.
3) There exists $\alpha > 0$ such that

$$\|\mathscr{A}u\|_{\mathbb{Y}} \geq \alpha \|u\|_{\mathbb{X}}, \quad \forall u \in \mathbb{X}.$$

4) There exists $\alpha > 0$ such that

$$\inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A}u, v)_{\mathbb{Y}}}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geq \alpha.$$

Proof. We only need to show 1) $\Leftrightarrow$ 2) as the equivalence between 2) and 3) is due to Lemma 15.1, and 4) is simply a restatement of 3). We have

$$\begin{array}{c} \mathscr{A}^* \text{ is surjective} \\ \Updownarrow \\ \mathsf{R}(\mathscr{A}^*) = \mathbb{X} \text{ and thus } \mathsf{R}(\mathscr{A}^*) \text{ closed} \\ \Updownarrow \\ \mathsf{N}(\mathscr{A}) = \mathsf{R}(\mathscr{A}^*)^\perp = \{\theta\} \text{ and } \mathsf{R}(\mathscr{A}) \text{ closed} \\ \Updownarrow \\ \mathscr{A} \text{ is injective and } \mathsf{R}(\mathscr{A}) \text{ closed.} \end{array}$$

The closed range Theorem 5.2

The following twin counterpart of Theorem 15.1 characterizes the surjectivity of $\mathscr{A}$ via the adjoint $\mathscr{A}^*$.

Theorem 15.2. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be linear and continuous. The following are equivalent:*

1) $\mathscr{A} : \mathbb{Y} \to \mathbb{X}$ is surjective.
2) $\mathscr{A}^*$ is injective and $\mathsf{R}(\mathscr{A}^*)$ is closed.
3) There exists $\alpha > 0$ such that

$$\|\mathscr{A}^* v\|_{\mathbb{X}} \geq \alpha \|v\|_{\mathbb{Y}}, \quad \forall v \in \mathbb{Y}.$$

4) There exists $\alpha > 0$ such that

$$\inf_{v \in \mathbb{Y}} \sup_{u \in \mathbb{X}} \frac{(u, \mathscr{A}^* v)_{\mathbb{X}}}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geq \alpha.$$

Combining Theorem 15.1 and Theorem 15.2 we see that $\mathscr{A}$ is bijective iff $\mathscr{A}^*$ is bijective. The more popular statement that leads to the Banach-Nečas-Babuška theorem for the variational equation is the following

Lemma 15.2. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be linear and continuous. The following are equivalent:*

1) $\mathscr{A}$ is bijective
2)
- *$\exists \alpha > 0$ such that $\inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A} u, v)_{\mathbb{Y}}}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geq \alpha$, and*
- *If $(\mathscr{A} u, v)_{\mathbb{Y}} = 0, \forall u \in \mathbb{X}$, then $v = \theta$.*

Proof. The proof is straightforward. Indeed, the first statement of 2) is the injectivity of $\mathscr{A}$ plus the closedness of $\mathsf{R}(\mathscr{A})$ due to Theorem 15.1, and the second statement of 2) can be written equivalently in terms of $\mathscr{A}^*$ as: "if $(u, \mathscr{A}^* v)_{\mathbb{Y}} = 0, \forall u \in \mathbb{X}$, then $v = \theta$", which is equivalent to "if $\mathscr{A}^* v = \theta$ then $v = \theta$", which in turn simply means $\mathsf{N}(\mathscr{A}^*) = \{\theta\}$, which then means $\mathscr{A}$ is surjective owing to Theorem 15.2.

State here that linear operator equation is equivalent to a matrix equation via matrix representation, and thus it is sufficient to consider a matrix example. Also after the matrix example, perhaps writing a remark on how the conditions are translated to the original linear operator equation

Example 15.1. Consider $\boldsymbol{A} : \mathbb{R}^n \mapsto \mathbb{R}^m$. We are interested in applying Lemma 15.2 to find conditions for the linear system of equations $\boldsymbol{Au} = \boldsymbol{y}$ to have a unique solution. To that end, we suppose $\boldsymbol{A}$ is bijective. Thus, both $\boldsymbol{A}$ and $\boldsymbol{A}^*$ are injective. By the rank-nullity theorem (8.5) we have

$$n = dim(\mathsf{N}(\boldsymbol{A})) + dim(\mathsf{R}(\boldsymbol{A})).$$

Since A is injective, we have $dim(\mathsf{N}(\boldsymbol{A})) = 0$, and hence

$$n = dim(\mathsf{R}(\boldsymbol{A})) \leq m.$$

Following similar arguments for the injectivity of $\boldsymbol{A}^*$, we have:

$$m = dim(\mathbb{R}(\boldsymbol{A}^*)) \leq n.$$

Therefore, it is necessary that $n = m$ for the bijectivity of $\boldsymbol{A}$. Further, the inf-sup condition in Lemma 15.2 says:

$$0 < \alpha \leq \inf_{\boldsymbol{u}\in\mathbb{R}^n} \sup_{\boldsymbol{v}\in\mathbb{R}^n} \frac{(\boldsymbol{Au}, \boldsymbol{v})}{\|\boldsymbol{u}\|_{\mathbb{R}^n} \|\boldsymbol{v}\|_{\mathbb{R}^n}} \leq \inf_{\boldsymbol{u}\in\mathbb{R}^n} \frac{\|\boldsymbol{Au}\|_{\mathbb{R}^n}}{\|\boldsymbol{u}\|_{\mathbb{R}^n}} = \sigma_{min}(\boldsymbol{A}).$$

$$\implies 0 < \alpha \leq \sigma_{min}(\boldsymbol{A}),$$

where $\sigma_{min}(\boldsymbol{A})$ denotes the smallest singular value of $\boldsymbol{A}$. We conclude the necessary and sufficient for a linear system of equations $\boldsymbol{Au} = \boldsymbol{y}$ to have a unique solution is that the matrix $\boldsymbol{A}$ is square and invertible. This is consistent with what we know from linear algebra.

Theorem 15.3 (Banach-Nečas-Babuška). *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be the unique linear and continuous operator (see Example 12.8) associated with a continuous sequilinear form $a : \mathbb{X} \times \mathbb{Y} \to \mathbb{F}$ such that*

$$(\mathscr{A}u, v)_{\mathbb{Y}} := a(u, v), \quad \forall u \in \mathbb{X} \text{ and } v \in \mathbb{Y},$$

where $|a(u, v)| \leq \beta \|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}$ and $0 < \beta < \infty$. The following are equivalent:

1) For all $y \in \mathbb{Y}$, there exists a unique solution $u \in \mathbb{X}$ such that

$$a(u, v) = (y, v)_{\mathbb{Y}}, \quad \forall v \in \mathbb{Y}.$$

2) There exists $\alpha > 0$ such that

C1) $\exists \alpha > 0$ *such that* $\inf_{u\in\mathbb{X}} \sup_{v\in\mathbb{Y}} \frac{a(u, v)}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geq \alpha$, *and*

C2) *If* $a(u,v) = 0, \forall u \in \mathbb{X}$, *then* $v = \theta$.

Furthermore, when either of the statements holds then the unique solution is stable in the following sense:

$$\|u\|_{\mathbb{Y}} \leq \frac{1}{\alpha} \|y\|_{\mathbb{Y}}.$$

Proof. The proof is obvious due to the definition $(\mathscr{A}u, v)_{\mathbb{Y}} := a(u,v)$, and thus the equivalent of the variational equation $a(u,v) = (y,v)_{\mathbb{Y}}$ and $\mathscr{A}u = y$. Specifically, owing to Lemma 15.2, statement 2) is equivalent to the bijectivity of $\mathscr{A}$. The stability of the solution u is the direction consequence of the boundedness from below of $\mathscr{A}$:

$$\alpha \|u\|_{\mathbb{X}} \leq \|\mathscr{A}u\|_{\mathbb{Y}} = \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A}u, v)_{\mathbb{Y}}}{\|v\|_{\mathbb{Y}}} = \sup_{v \in \mathbb{Y}} \frac{a(u,v)_{\mathbb{Y}}}{\|v\|_{\mathbb{Y}}} = \sup_{v \in \mathbb{Y}} \frac{(y,v)_{\mathbb{Y}}}{\|v\|_{\mathbb{Y}}} = \|y\|_{\mathbb{Y}}.$$

Remark 15.1. The condition $\inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{a(u,v)}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geq \alpha$, is known as the inf-sup condition, and, as we have shown, it is nothing more than the restatement of the boundedness from below of the associated linear operator $\mathscr{A}$ or equivalently the injectivity of $\mathscr{A}$ plus its closed range.

Example 15.2. Let us continue our study of the differentiable equation in the weak setting developed in Example 13.11. Consider solving the diferential equation $u' := \mathscr{D}u = f$ in $(0,1)$ with $u(0) = 0$ and $f \in \mathbb{L}^2(0,1)$. The corresponding weak form of the problem is formulated as: seek $u \in \mathbb{H}^1_0(0,1) := \{u \in \mathbb{L}^2(0,1),\ \ u' \in \mathbb{L}^2(0,1),\ u(0) = 0\}$ such that:

$$(u', v)_{\mathbb{L}^2} = (f, v)_{\mathbb{L}^2}, \quad \forall v \in \mathbb{L}^2(0,1).$$

We choose $\mathbb{X} = \mathbb{H}^1_0(0,1)$, $\mathbb{Y} = \mathbb{L}^2(0,1)$ and $\mathbb{F} = \mathbb{R}$, and are going to use Theorem 15.3 to show that the differential equation is well-posed.

The continuity of the bilinear form $a(u,v)$ is clear as

$$|a(u,v)| = |(u', v)| \leq \|u'\|_{\mathbb{L}^2} \|v\|_{\mathbb{L}^2} \leq \|u\|_{\mathbb{H}^1} \|v\|_{\mathbb{L}^2}.$$

We next verify the inf-sup condition Item **C1)**. By a simple integration and the Cauchy-Schwarz inequality (see also Example 16.1) we obtain the following the Poincaré-Friedrichs inequality

$$\|u'\|_{\mathbb{L}^2} \geq \ \|u\|_{\mathbb{L}^2},$$

and thus the inf-sup condition holds since

$$\inf_{u \in \mathbb{H}^1} \sup_{v \in \mathbb{L}^2} \frac{a(u,v)}{\|u\|_{\mathbb{H}^1} \|v\|_{\mathbb{L}^2}} = \inf_{u \in \mathbb{H}^1} \frac{\|u'\|_{\mathbb{L}^2}}{\|u\|_{\mathbb{H}^1}} \geq \frac{1}{\sqrt{2}}.$$

Now we verify the injectivity of the adjoint, i.e Item **C2)**. We start from

$$(u', v)_{\mathbb{L}^2} = 0, \quad \forall u \in \mathbb{H}_0^1(0,1). \tag{15.1}$$

Since $\mathcal{C}_0^\infty(0,1) \subset \mathbb{H}_0^1(0,1)$, we have:

$$(\psi', v)_{\mathbb{L}^2} = 0, \quad \forall \psi \in \mathcal{C}_0^\infty(0,1).$$

By definition of the distributional derivative we arrive at

$$\langle \psi, v' \rangle_{\mathbb{L}^2} = 0, \quad \forall \psi \in \mathcal{C}_0^\infty(0,1),$$

which implies that v is a constant function. Now in (15.1) taking $u = x$ we have

$$\int_0^1 v dx = 0,$$

which means $v = 0$. Thus, the differential equation is well-posed with the setting $\mathbb{X} = \mathbb{H}_0^1(0,1)$, $\mathbb{Y} = \mathbb{L}^2(0,1)$.

Example 15.3 (Friedrichs' system). We consider the abstract problem $\mathscr{A}u = y$ where $\mathscr{A}$ is the Friedrichs' operator defined in Example 12.10. We choose $\mathbb{Y} = \mathbb{L}^2(\Omega)$ and

$$\mathbb{X} = \{\boldsymbol{u} \in \mathbb{H}_{\mathscr{A}} : (\boldsymbol{D} - \boldsymbol{M})\boldsymbol{u} = \boldsymbol{0} \text{ on } \partial\Omega\},$$

with the inner product $(u, w)_{\mathbb{X}} := (u, w)_{\mathbb{Y}} + (\mathscr{B}u, \mathscr{B}w)_{\mathbb{Y}}$, and hence the induced graph norm $\|u\|_{\mathbb{X}} = \sqrt{\|u\|_{\mathbb{Y}}^2 + \|\mathscr{B}u\|_{\mathbb{Y}}^2}$. We can show that $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ is bijective (see [51, Theorem 5.7]), and thus applying Theorem 15.3 shows that $\mathscr{A}u = y$ is well-posed for any $y \in \mathbb{Y}$. The beauty here is that this single proof is applicable for a large class of PDEs [58, 52, 78, 54].

Need to provide the proof that the inf-sup constants for both the primal and the adjoint problem are the same.

When $\mathbb{Y} = \mathbb{X}$ and the sequilinear form is symmetric, the inf-sup condition is both necessary and sufficient for bijectivity.

Lemma 15.3. *Consider the variational equation $a(u, v) = (y, v)_{\mathbb{X}}$, $\forall v \in \mathbb{X}$, with a continuous sequilinear form $a : \mathbb{X} \times \mathbb{X} \to \mathbb{F}$. Suppose $a(\cdot, \cdot)$ is symmetric, i.e., $a(w, v) = \overline{a(v, w)}$. Then, there exists a unique solution u iff*

$$\exists \alpha > 0 \text{ such that } \inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{a(u, v)}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geq \alpha.$$

Proof. We need to prove only the second condition in the second statement of Theorem 15.3, namely the injectivity of the adjoint in Item **C2)**. But this is obvious due to symmetry:

$$0 = \overline{a(w,v)} = a(v,w), \quad \forall w \in \mathbb{X} \implies 0 = \sup_{w\in\mathbb{X}} \frac{a(v,w)}{\|w\|_{\mathbb{X}}} \geq \alpha \|v\|_{\mathbb{X}} \implies v = \theta.$$

Remark 15.2. Note that the symmetry of $a(w,v)$ is equivalent to the self-adjointness of its associate linear and continuous operator $\mathscr{A}$ defined in Theorem 15.3. Indeed, since $a : \mathbb{X} \times \mathbb{X} \to \mathbb{F}$, we have

$$(\mathscr{A}w, v)_{\mathbb{X}} = a(w,v) = \overline{a(v,w)} = \overline{(\mathscr{A}v, w)}_{\mathbb{X}} = (w, \mathscr{A}v)_{\mathbb{X}},$$

which means $\mathscr{A}^* = \mathscr{A}$.

On the other hand, when $\mathbb{Y} = \mathbb{X}$ and the sequilinear form is coercive, the condition for bijectivity is simpler.

Lemma 15.4 (The Lax-Milgram lemma). *Consider the variational equation $a(u,v) = (y,v)_{\mathbb{X}}$, $\forall v \in \mathbb{X}$, with a continuous sequilinear form $a : \mathbb{X}\times\mathbb{X} \to \mathbb{F}$. If*

$$a(v,v) \geq \alpha \|v\|_{\mathbb{X}}^2, \quad \forall v \in \mathbb{X}, \tag{Coercivity}$$

then there exists a unique solution u and $\|u\| \leq \frac{1}{\alpha}\|y\|_{\mathbb{X}}$.

Proof. We need to verify the two conditions in the second statement of Theorem 15.3. The inf-sup condition in Item **C1)** is clear from (Coercivity) as

$$\alpha \|v\|_{\mathbb{X}} \leq \frac{a(v,v)}{\|v\|_{\mathbb{X}}} \leq \sup_{w\in\mathbb{X}} \frac{a(v,w)}{\|w\|_{\mathbb{X}}}.$$

For the injectivity of the adjoint in Item **C2)**, we note that

$$a(w,v) = 0, \quad \forall w \in \mathbb{X} \implies 0 = \sup_{w\in\mathbb{X}} a(w,v) \geq a(v,v) \geq \alpha \|v\|_{\mathbb{X}}^2 \implies v = \theta.$$

When the sequilinear form is symmetric and positive, it turns out that (Coercivity) is both sufficient and necessary.

Corollary 15.1. *Consider the variational equation $a(u,v) = (y,v)_{\mathbb{X}}$, $\forall v \in \mathbb{X}$, with a continuous sequilinear form $a : \mathbb{X}\times\mathbb{X} \to \mathbb{F}$, with continuity constant β. Suppose $a(\cdot,\cdot)$ is symmetric, i.e., $a(w,v) = \overline{a(v,w)}$ and positive, i.e., $a(v,v) > 0, \forall v \neq \theta$. Then, there exists a unique solution u iff the coercivity condition (Coercivity) holds.*

Proof. Lemma 15.4 proves the sufficiency, and we need to show the necessity. Since the sequilinear form $a(\cdot,\cdot)$ is symmetric and positive, it defines an inner product in $\mathbb{X}$ and the induced norm is

$$\|v\|_a := \sqrt{a(v,v)}.$$

Thus, by the inf-sup condition, the Cauchy-Schwarz inequality, and the continuity of $a(\cdot,\cdot)$, we have

$$\alpha \|v\|_{\mathbb{X}} \leq \sup_{w \in \mathbb{X}} \frac{a(w,v)}{\|w\|_{\mathbb{X}}} \leq \sup_{w \in \mathbb{X}} \frac{\sqrt{a(v,v)}\sqrt{a(w,w)}}{\|w\|_{\mathbb{X}}} \leq \sqrt{\beta}\sqrt{a(v,v)},$$

and this ends the proof.

Example 15.4. We now consider the variational equation $a(u,v) = (y,v)_{\mathbb{Y}}$ where the bilinear form $a(u,v)$, the spaces $\mathbb{X}$ and $\mathbb{Y}$, and other specifications are described in Example 12.8. We have shown that $a(u,v)$ is symmetric and continuous on $\mathbb{H}_0^1(\Omega)$. It is clearly positive if we assume that z is bounded away from $-\infty$. What remains is to show that $a(u,v)$ is coercive. Recall the Poincaré-Friedrichs inequality (see, e.g., [51, 11, 28]) for $\mathbb{H}_0^1(\Omega)$: there exists a constant c depending on only Ω such that

$$c\|u\|_{\mathbb{L}^2} \leq \|\nabla u\|_{\mathbb{L}^2}, \quad \forall u \in \mathbb{H}_0^1(\Omega).$$

It follows that for any $v \in \mathbb{H}_0^1(\Omega)$ we have

$$a(v,v) = (e^z \nabla v, \nabla v)_{\mathbb{L}^2}^2 \geq e^{\inf z} \|\nabla v\|_{\mathbb{L}^2}^2 \geq \min\{1,c\} \frac{1}{2} e^{\inf z} \|v\|_{\mathbb{H}^1}^2,$$

and this ends the proof.

15.1 Appendix

Proof (of Lemma 15.1). For the necessary, the injectivity is clear. Now, let $\{y_i\}_{i=1}^{\infty} \subset \mathsf{R}(\mathscr{A})$ and $y_i \xrightarrow{\mathbb{Y}} y$ and we need to show that $y \in \mathsf{R}(\mathscr{A})$. There exists $\{x_i\}_{i=1}^{\infty} \subset \mathbb{X} : y_i = \mathscr{A} x_i$. For any $\varepsilon > 0$, there exists an integer $n = n(\varepsilon)$ such that for all $i, j > n$ we have

$$\begin{array}{l} \alpha \|x_i - x_j\|_{\mathbb{X}} \leq \|\mathscr{A} x_i - \mathscr{A} x_j\|_{\mathbb{Y}} \|y_i - y_j\|_{\mathbb{Y}} < \varepsilon \\ \qquad\qquad \Downarrow \\ \{x_i\}_{i=1}^{\infty} \text{ is Cauchy} \implies x_i \xrightarrow{\mathbb{X}} x \\ \qquad\qquad \Downarrow \qquad\qquad\qquad\qquad \text{continuity of } \mathscr{A} \\ y \xleftarrow{\mathbb{Y}} y_i = \mathscr{A} x_i \xrightarrow{\mathbb{Y}} \mathscr{A} x, \end{array}$$

and thus $\mathsf{R}(\mathscr{A})$ is closed.

For the sufficiency, $\mathsf{R}(\mathscr{A})$ is a Banach space due to its closedness. $\mathscr{A} : \mathbb{X} \to \mathsf{R}(\mathscr{A})$ is thus bijective, which in turn implies $\mathscr{A}^{-1}$ is linear and continuous owing to the Open Mapping Theorem [28, 112, 11, 127]. Thus, let $y = \mathscr{A} x$, there exists $\beta > 0$ such that $\left\|\mathscr{A}^{-1} y\right\|_{\mathbb{X}} \leq \beta \|y\|_{\mathbb{Y}} \implies \|x\|_{\mathbb{X}} \leq \beta \|\mathscr{A} x\|_{\mathbb{Y}}$, and hence $\mathscr{A}$ is bounded below.

Chapter 16
Understanding Sturm-Liouville problem and generalized Fourier series using adjoint

Abstract

In this section, we are interested in self-adjoint operator $\mathscr{A}$ in infinite dimensions. Our exposition mostly follows [132]. To the end of this section, for any $\mathscr{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ we assume that its domain $\mathsf{D}(\mathscr{A}) \subseteq \mathbb{X}$ is dense in $\mathbb{X}$.

Definition 16.1 (Closed linear operators). A linear operator $\mathscr{A} \in \mathcal{L}(\mathbb{X}, \mathbb{Y})$ is called closed iff its graph

$$\mathcal{G}_{\mathscr{A}} := \{[u, \mathscr{A}u] : u \in \mathscr{D}(\mathscr{A})\}$$

is closed in the product topology on $\mathbb{X} \times \mathbb{Y}$.

Clearly, any continuous linear operator is necessarily closed. The following are standard results for closed operators [132].

Lemma 16.1. *The following hold:*

- *If $\mathscr{A}$ is closed, so is $\mathscr{A}^*$.*
- *$\mathscr{A}$ is closed and $\mathsf{D}(\mathscr{A}) = \mathbb{X}$ iff $\mathscr{A} \in \mathscr{B}(\mathbb{X}, \mathbb{Y})$.*
- *If $\mathsf{D}(\mathscr{A}) = \mathbb{X}$, then $\mathscr{A}^*$ is continuous, and hence $\mathsf{D}(\mathscr{A}^*)$ is closed.*
- *If $\mathscr{A}$ is closed, then $\mathsf{D}(\mathscr{A}^*)$ is dense in $\mathbb{Y}$.*

We consider operators with $\mathbb{X} = \mathbb{Y}$ in this section. Suppose $\mathbb{V}$ is dense in $\mathbb{X}$ and the injection $\mathbb{V} \to \mathbb{X}$ is compact, and for simplicity in writing we denote $\mathbb{V} \xhookrightarrow[\text{dense}]{\text{compact}} \mathbb{X}$. We shall limit ourself to linear operator $\mathscr{A} : \mathsf{D}(\mathscr{A}) \subset \mathbb{V} \xhookrightarrow[\text{dense}]{\text{compact}} \mathbb{X} \to \mathbb{X}$ with the domain defined as

$$\mathsf{D}(\mathscr{A}) := \{x \in \mathbb{V} : \mathscr{A}x \in \mathbb{X}\}.$$

We also assume the following.

1. The associate sequilinear form $a(u, v) := (\mathscr{A}u, v)_{\mathbb{X}}$ is defined for $u \in \mathsf{D}(\mathscr{A})$ and $v \in \mathbb{V}$, and is continuous in v on $\mathbb{V}$ with respect to the $\mathbb{X}$-norm topology.

2. The $a(u,v)$ is $\mathbb{V}$-elliptic in the following sense: there exists $c > 0$ such that
$$a(v,v) + \overline{a(v,v)} \geq 2c \|v\|_{\mathbb{V}}^2 .$$
3. Finally, $a(\cdot,\cdot)$ is symmetric, i.e.,
$$a(u,v) = \overline{a(v,u)}, \quad \forall u, v \in \mathbb{V},$$
which is, again, equivalent to the self-adjointness of $\mathscr{A}$ on $\mathbb{V}$.

Recall that Lemma 7.1 holds in this case: in particular, eigenvalues of $\mathscr{A}$ are real and its eigenfunctions corresponding to distinct eigenvalues are orthogonal to each other. The following theorem provides further characteristics of eigenpairs of $\mathscr{A}$.

Theorem 16.1. *Suppose all the aforementioned assumptions hold for $\mathscr{A}$. Then, there is a countable sequence of eigenpairs $\{\lambda_n, v_n\}_{n=1}^{\infty}$ such that*

- $\mathscr{A} v_n = \lambda_n v_n$, *where* $\|v\|_{\mathbb{X}} = 1$,
- $(v_n, v_m)_{\mathbb{X}} = 0$ *for all* $n \neq m$,
- $0 < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \to \infty$ *when* $n \to \infty$, *and*
- $\{v_n\}_{n=1}^{\infty}$ *is an orthonormal basis of* $\mathbb{X}$.

Proof. We sketch the proof and more details can be found in [131]. The keys are: i) the $\mathbb{V}$-elliptic condition implies coercivity of $\mathscr{A}$ on $\mathbb{V}$. By Lax-Milgram Lemma 15.4, $\mathscr{A} : \mathbb{V} \to \mathbb{V}' \equiv \mathbb{V}$ is a continuous bijection. The restriction of $\mathscr{A}$ on $\mathsf{D}(\mathscr{A})$ is thus surjective on $\mathbb{X}$, and $\mathscr{A}^{-1} : \mathbb{X} \to \mathsf{D}(\mathscr{A}) \subset \mathbb{V}$ exists and is continuous by the coercivity; ii) the compact injection of $\mathbb{V}$ in $\mathbb{X}$ then implies that $\mathscr{A}^{-1} : \mathbb{X} \to \mathbb{X}$ is compact. Hilbert-Schmidt Theorem 14.1 then ensures the existence of the eigenpairs μ_n, v_n of $\mathscr{A}^{-1}$, where $\{v_n\}_n$ is a basis of $\mathsf{D}(\mathscr{A})$, and hence the eigenpairs λ_i, v_n of $\mathscr{A}$ where $\lambda_n = 1/\mu_n$; and iii) the $\mathbb{V}$-elliptic condition also implies $\mathscr{A}$ is a closed operator and this leads to the conclusion that $\{v_n\}_{n=1}^{\infty}$ is a also basis of $\mathbb{X}$.

Theorem 16.1 provides a generalized Fourier series in $\mathbb{X}$. In particular, for any function $f \in \mathbb{X}$, we have
$$f = \sum_{n=1}^{\infty} (v_n, f)_{\mathbb{X}} \, v_n, \tag{16.1}$$
where equality means convergence in the topology generated by the $\mathbb{X}$-norm. More generally, we have the following resolution of identity in $\mathbb{X}$.

Corollary 16.1 (Resolution of identity). *Suppose the setting, and hence all the results, in Theorem 16.1 holds. Then*
$$\mathscr{I} = \sum_{n=1}^{\infty} (v_n, \cdot)_{\mathbb{X}} \, v_n,$$

where $\mathscr{I}$ is the identity operator and the convergence of the series is in the strong sense, that is, the series in (16.1) converges in norm topology of $\mathbb{X}$.

Proof. Though there is nothing to prove here, we shall show that the series is an orthogonal projection onto $\mathbb{X}$. To that end, to avoid confusion, let us define

$$\mathscr{P} := \sum_{n=1}^{\infty} (v_n, \cdot)_{\mathbb{X}} \, v_n.$$

First, we show that $\mathscr{P}^2 = \mathscr{P}$. Indeed, due to the continuity of the inner product, we have

$$\mathscr{P}^2 f = \mathscr{P}(\mathscr{P} f) = \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} (v_m, f)_{\mathbb{X}} (v_n, v_m)_{\mathbb{X}} \, v_n = \sum_{n=1}^{\infty} (v_n, f)_{\mathbb{X}} \, v_n = \mathscr{P} f.$$

Second, $\mathscr{P}$ is self-adjoint. Indeed,

$$\begin{aligned} (\mathscr{P} f, g)_{\mathbb{X}} = \left(\sum_{n=1}^{\infty} (v_n, f)_{\mathbb{X}} \, v_n, g \right)_{\mathbb{X}} &= \sum_{n=1}^{\infty} (f, v_n)_{\mathbb{X}} (v_n, g)_{\mathbb{X}} \\ &= \left(f, \sum_{n=1}^{\infty} (v_n, g)_{\mathbb{X}} \, v_n \right)_{\mathbb{X}} = (f, \mathscr{P} g)_{\mathbb{X}}, \end{aligned}$$

where we have used the continuity of the inner product.

Below are a few special cases that lead to the standard Fourier series and a general Sturm-Liouville problem.

Example 16.1. Consider the following Sturm-Liouville problem: seek λ and v such that

$$\begin{cases} -\dfrac{d^2 v}{dx^2} = \lambda v & \text{in } \Omega = (0,1), \\ v = 0 & \text{on } \partial\Omega = \{0,1\}, \end{cases} \tag{16.2}$$

where the derivative is understood in the classical sense. We solve this problem using Theorem 16.1. To that end, we define the symmetric sequilinear form as

$$a(u,v) := \left(\frac{du}{dx}, \frac{dv}{dx} \right)_{\mathbb{L}^2(\Omega)} = \int_0^1 \overline{\frac{du}{dx}} \frac{dv}{dx} \, dx,$$

for all $u, v \in \mathbb{V}$ with $\mathbb{V} := \mathbb{H}_0^1(0,1)$ and thus the derivatives in $a(\cdot,\cdot)$ are understood in the weak sense. Note that $\mathbb{V} \xhookrightarrow[\text{dense}]{\text{compact}} \mathbb{X} := \mathbb{L}^2(0,1)$ [132], and the continuity on $\mathbb{V}$ of $a(\cdot,\cdot)$ is straightforward. By the fundamental theorem of calculus and Cauchy-Schwarz inquality we can easily arrive at a Poincaré-Friedrichs inequality: $\forall v \in \mathcal{C}_0^1[0,1]$

$$\left\| \frac{dv}{dx} \right\|_{\mathbb{L}^2(0,1)} \geq \|v\|_{\mathbb{L}^2(0,1)},$$

which also holds for any $v \in \mathbb{H}_0^1(0,1)$ due to the density of $\mathcal{C}_0^1(0,1)$ in $\mathbb{V}$. This leads to the $\mathbb{V}$-ellipticity of $a(u,v)$ as

$$a(v,v) + \overline{a(v,v)} \geq \|v\|^2_{\mathbb{H}_0^1(0,1)}.$$

Thus, Theorem 16.1 ensures that there is a complete orthonormal eigenfunctions of $\mathscr{A} : \mathsf{D}(\mathscr{A}) \to \mathbb{L}^2(0,1)$ in $\mathbb{L}^2(0,1)$, where

$$\mathsf{D}(\mathscr{A}) := \left\{ w \in \mathbb{H}_0^1(0,1) : \mathscr{A}w \in \mathbb{L}^2(0,1) \right\},$$

and $(\mathscr{A}u, v)_{\mathbb{L}^2(0,1)} := a(u,v)$ for all $u \in \mathsf{D}(\mathscr{A})$ and $v \in \mathbb{H}_0^1(0,1)$. Let us determine what $\mathsf{D}(\mathscr{A})$ is. We have, by definition of $a(\cdot,\cdot)$,

$$\begin{aligned}
(\mathscr{A}w, v)_{\mathbb{L}^2(0,1)} &= \int_0^1 \overline{\frac{dw}{dx}} \frac{dv}{dx}\, dx && \forall v \in \mathbb{H}_0^1(0,1) \\
&\Updownarrow && \text{density of } \mathcal{C}_0^\infty(0,1) \text{ in } \mathbb{H}_0^1(0,1) \\
(\mathscr{A}w, \varphi)_{\mathbb{L}^2(0,1)} &= \int_0^1 \overline{\frac{dw}{dx}} \frac{d\varphi}{dx}\, dx && \forall \varphi \in \mathcal{C}_0^\infty(0,1) \\
&\Updownarrow && \text{definition of distributional derivative} \\
(\mathscr{A}w, \varphi)_{\mathbb{L}^2(0,1)} &= \left\langle -\frac{d^2w}{dx^2}, \varphi \right\rangle && \forall \varphi \in \mathcal{C}_0^\infty(0,1) \\
&\Updownarrow \\
\mathscr{A}w &= -\frac{d^2w}{dx^2} && \text{in } \mathbb{L}^2(0,1)
\end{aligned}$$

As a result, $\mathsf{D}(\mathscr{A}) = \mathbb{H}_0^2(0,1)$.

Each eigenpair λ_n and $v_n \in \mathsf{D}(\mathscr{A})$ satisfies

$$\mathscr{A}v_n = \lambda_n v_n \text{ in } \mathbb{L}^2(0,1),$$

i.e,

$$-\frac{d^2v_n}{dx^2} = \lambda_n v_n, \text{ in } \mathbb{L}^2(0,1), \tag{16.3}$$

which is equivalent to

$$v_n = -\int\int \lambda_n v_n \, dx\, dy,$$

Owing $v_n \in \mathsf{D}(\mathscr{A})$, and the embedding of $\mathbb{H}_0^1(0,1)$ in $\mathcal{C}_0(0,1)$ (see, e.g., [28, Theorem 8.2]) the eigenvalue problem (16.3) holds in the classical sense, which is exactly the Sturm-Liouville problem (16.2). Thus, eigenfunctions of the Sturm-Liouville problem (16.2) forms a complete basis for $\mathbb{L}^2(0,1)$. By a

simple integrations, we have $v_n = \sqrt{2}\sin(n\pi x)$, $n = 1, 2, \ldots$, which is exactly a Fourier basis (*sine series*) for $\mathbb{L}^2(0,1)$.

Corollary 16.2. *Let $\mathbb{V}$ and $\mathbb{X}$ be given in Theorem 16.1. Suppose the sequilinear form $a(\cdot,\cdot)$ is symmetric and continuous on $\mathbb{V}$. Assume that there exists $c > 0$ such that for some $\lambda \in \mathbb{R}$*

$$a(v,v) + \overline{a(v,v)} + 2\lambda \|v\|_{\mathbb{X}}^2 \geq 2c\|v\|_{\mathbb{V}}^2, \quad \forall v \in \mathbb{V}.$$

Then, there exists an orthonormal sequence of eigenfunctions of $\mathscr{A}$, which is a basis for $\mathbb{X}$ and the corresponding eigenvalues satisfies $-\lambda < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \to \infty$ when $n \to \infty$.

Proof. Let us define the sequilinear $b(u,v) := a(u,v) + \lambda(u,v)_{\mathbb{X}}$. The continuity and symmetry of $b(\cdot,\cdot)$ are clear. The linear operator $\mathscr{B}$ associated with $b(u,v)$ is given by $\mathscr{B}u = \mathscr{A}u + \lambda u$ for any $u \in \mathsf{D}(\mathscr{B}) \equiv \mathsf{D}(\mathscr{A})$, and $b(u,v) = (\mathscr{B}u, v)_{\mathbb{X}}$ for all $u \in \mathsf{D}(\mathscr{B})$ and $v \in \mathbb{V}$. Furthermore, we have

$$b(v,v) + \overline{b(v,v)} \geq 2c\|v\|_{\mathbb{V}}^2, \quad \forall v \in \mathbb{V},$$

i.e., $b(u,v)$ is $\mathbb{V}$-elliptic. Theorem 16.1 thus applies to $b(\cdot,\cdot)$. In particular, there exist eigenpairs $\{\gamma_n, v_n\}_{n=1}^{\infty}$ of $\mathscr{B}$ such that $\mathscr{A}v_n + \lambda v_n = \mathscr{B}v_n = \gamma_n v_n$ and $0 < \gamma_1 \leq \gamma_2 \leq \ldots \leq \gamma_n \to \infty$ when $n \to \infty$. As a result, $\{\lambda_n, v_n\}_{n=1}^{\infty}$ with $\lambda_n := \gamma_n - \lambda$, and thus $-\lambda < \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n \to \infty$ when $n \to \infty$, and $\{v_n\}_{n=1}^{\infty}$ being a basis of $\mathbb{X}$, and this concludes the proof.

Remark 16.1. We can use Corollary 16.2 for Example 16.1 as well. In particular, we can choose $\mathbb{V} = \mathbb{H}^1(0,1)$, and $b(u,v) := a(u,v) + \lambda(u,v)_{\mathbb{L}^2(0,1)}$ is coercive on $\mathbb{V}$ for any $\lambda > 0$ with $c = \min\{1, \lambda\}$. The rest of the arguments are similar and the same results are obtained.

Example 16.2. Consider the following Sturm-Liouville problem: seek κ and v such that

$$\begin{cases} -\dfrac{d^2v}{dx^2} = \kappa v & \text{in } \Omega = (0,1), \\ \dfrac{dv}{dx} = 0 & \text{on } \partial\Omega = \{0,1\}, \end{cases} \tag{16.4}$$

where the derivative is understood in the classical sense. The sequilinear $a(\cdot,\cdot)$ is defined the same as in Example 16.1 with $\mathbb{V} := \mathbb{H}^1(0,1)$ and $\mathbb{X} := \mathbb{L}^2(0,1)$. We still have $\mathbb{V} \xhookrightarrow[\text{dense}]{\text{compact}} \mathbb{X} := \mathbb{L}^2(0,1)$ [132]. The linear operator $\mathscr{A}$ associated with $a(\cdot,\cdot)$ defined via the identity $(\mathscr{A}u, v)_{\mathbb{X}} = a(u,v)$ for all $u \in \mathsf{D}(\mathscr{A})$ where

$$\mathsf{D}(\mathscr{A}) := \left\{ w \in \mathbb{V} : \mathscr{A}w \in \mathbb{X} \text{ and } \frac{dw}{dx} = 0 \text{ on } \partial\Omega = \{0,1\} \right\},$$

and $v \in \mathbb{V}$. Note that unlike Example 16.1 in which the boundary conditions are naturally incorporated in $\mathbb{V}$, we have to build the boundary conditions

in the definition of the domain of $\mathscr{A}$ in order to associate $a(\cdot,\cdot)$ with the eigenvalue problem (16.4). The continuity and symmetry of $a(\cdot,\cdot)$ on $\mathbb{V}$ are clear. Furthermore, $a(\cdot,\cdot)$ satisfies Corollary 16.2 for any $\lambda > 0$ and $c = \min\{1,\lambda\}$. Using a similar argument as in Example 16.1, the eigenpairs λ_n, v_n of $\mathscr{A}$ are exactly the solutions of (16.4), and in particular $v_n = \sqrt{2}\cos(n\pi x)$ for $n \geq 1$ and $v_0 = 1$. By Corollary 16.2, $\{v_n\}_{n=0}^{\infty}$ is another Fourier basis (*cosine series*) for $\mathbb{L}^2(0,1)$.

Example 16.3 (A more general Sturm-Liouville problem). We now generalize Example 16.1 and Example 16.2: seek κ and v such that

$$\begin{cases} \dfrac{1}{\rho}\left[\dfrac{d}{dx}\left(p\dfrac{dv}{dx}\right)+qv\right] = \kappa v & \text{in } \Omega=(0,1), \\ \alpha v(0)+\beta\dfrac{dv}{dx}(0)=0, & \gamma v(1)+\delta\dfrac{dv}{dx}(1)=0, \end{cases} \tag{16.5}$$

where $\rho \in \mathcal{C}[0,1]$ and $\rho > 0$, $p \in \mathcal{C}^1[0,1]$ and $p < 0$, and $q \in \mathcal{C}[0,1]$. Here, the constants α,β,γ, and δ satisfy $\alpha^2+\beta^2 \neq 0$, and $\gamma^2+\delta^2 \neq 0$. We take $\mathbb{X} := \mathbb{L}^2_\rho(0,1)$ is the $\mathbb{L}^2(0,1)$ space with the weighted inner product: $(u,v)_{\mathbb{L}^2(0,1),\rho} := (\rho u, v)_{\mathbb{L}^2(0,1)}$. We choose the sequilinear form $a(\cdot,\cdot)$ as

$$a(u,v) := \left(p\frac{du}{dx},\frac{dv}{dx}\right)_{\mathbb{L}^2(0,1)} + (qu,v)_{\mathbb{L}^2(0,1)},$$

with the derivative understood in the weak sense. By taking $\mathbb{V} = \mathbb{H}^1_\rho(0,1)$, where $\mathbb{H}^1_\rho(0,1)$ is $\mathbb{H}^1(0,1)$ based on $\mathbb{L}^2_\rho(0,1)$, it is clear that $a(\cdot,\cdot)$ is continuous and symmetric on $\mathbb{V}$. The linear operator $\mathscr{A}$ associated with $a(\cdot,\cdot)$ is defined via the identity $(\mathscr{A}u,v)_{\mathbb{X}} = a(u,v)$ for all $u \in \mathsf{D}(\mathscr{A})$ where

$$\mathsf{D}(\mathscr{A}) := \left\{ w \in \mathbb{H}^1_\rho(0,1) : \mathscr{A}w \in \mathbb{X}, \alpha w(0)+\beta\frac{dw}{dx}(0)=0 \text{ and } \gamma w(1)+\delta\frac{dw}{dx}(1)=0 \right\},$$

and $v \in \mathbb{V}$. By a similar distributional argument as in Example 16.1, one can show that $\mathscr{A}w = \dfrac{1}{\rho}\left[\dfrac{d}{dx}\left(p\dfrac{dw}{dx}\right)+qw\right]$ in $\mathbb{L}^2(0,1)$. Furthermore, it is clear that $a(\cdot,\cdot)$ satisfies Corollary 16.2 for any $\lambda > -q$ and $c = \min\left\{\left\|\dfrac{\rho}{p}\right\|_\infty^{-1}, \left\|\dfrac{\rho}{\lambda+q}\right\|_\infty^{-1}\right\}$. We thus conclude that the eigenpairs λ_n and $v_n \in \mathsf{D}(\mathscr{A})$ satisfy

$$\frac{1}{\rho}\left[\frac{d}{dx}\left(p\frac{dv_n}{dx}\right)+qv_n\right] = \lambda_n v_n, \tag{16.6}$$

which is equivalent to

$$v_n = \int \frac{1}{p}\int \rho(\lambda_n - q)\, v_n\, dx\, dy,$$

which, together with the embedding result in [28, Theorem 8.2], shows that the eigenvalue problem (16.6) holds in the classical sense, which is exactly the Sturm-Liouville problem (16.5). Thus, eigenfunctions of the Sturm-Liouville problem (16.5) forms a complete basis for $\mathbb{L}^2(0,1)$. However, it is not so clear how to calculate the eigenpairs analytically as we have done for the previous examples.

Example 16.4. Next, we consider the following Sturm-Liouville problem that is not covered by Example 16.3: seek κ and v such that

$$\begin{cases} -\dfrac{d^2v}{dx^2} = \kappa v & \text{in } \Omega = (0,1), \\ v(0) = v(1) & \text{and } \dfrac{dv}{dx}(0) = \dfrac{dv}{dx}(1), \end{cases} \tag{16.7}$$

where the derivative is understood in the classical sense. The sequilinear $a(\cdot,\cdot)$ is defined the same as in Example 16.1 with $\mathbb{V} := \mathbb{H}^1(0,1)$ and $\mathbb{X} := \mathbb{L}^2(0,1)$. The linear operator $\mathscr{A}$ associated with $a(\cdot,\cdot)$ defined via the identity $(\mathscr{A}u, v)_{\mathbb{X}} = a(u,v)$ for all $u \in \mathsf{D}(\mathscr{A})$, where

$$\mathsf{D}(\mathscr{A}) := \left\{ w \in \mathbb{V} : \mathscr{A}w \in \mathbb{X}, v(0) = v(1) \text{ and } \frac{dv}{dx}(0) = \frac{dv}{dx}(1) \right\},$$

and $v \in \mathbb{V}$. Similar to Example 16.2, $a(\cdot,\cdot)$ is obviously symmetric and continuous on $\mathbb{V}$. Furthermore, $a(\cdot,\cdot)$ satisfies Corollary 16.2 for any $\lambda > 0$ and $c = \min\{1, \lambda\}$. Using a similar distributional argument as in Example 16.1, the eigenpairs λ_n, v_n of $\mathscr{A}$ are exactly the solutions of (16.7), and in particular $v_0 = 1$ and $v_n \in \left\{ \sqrt{2}\cos(2n\pi x), \sqrt{2}\sin(2n\pi x : n \in \mathbb{N}) \right\}$, which is the usual Fourier basis for $\mathbb{L}^2(0,1)$.

From the above examples, a few observations are in order. First, the view of (generalized) Fourier series from self-adjoint Sturm-Liouville operators immediately provides a rigorous convergence guarantee for the (generalized) Fourier series in the $\mathbb{L}^2$-sense. This view also shows that there are other orthogonal bases for $\mathbb{L}^2(a,b)$ and we can in principle find them by solving the corresponding Sturm-Liouville eigenvalue problems. Second, the results also show that $\mathbb{L}^2(a,b)$ is a separable Hilbert space. Third, the results are not restricted to $\mathbb{L}^2$ spaces over compact subsets in $\mathbb{R}$ but are also valid for any compact subsets in $\mathbb{R}^n$ using tensor product, dilation, and translation (see, e.g., [11]).

Chapter 17
Efficient PDE-constrained optimization with Adjoint

Abstract We have seen the important role of adjoint in constrained optimization in Chapter 9, especially constrained optimization with equality constraints that have separable structure (see Corollary 9.2 and Lemma 9.4). We have also seen what the adjoint looks like and how it helps compute the gradient of deep neural networks (DNN) efficiently as backpropagation in Chapter 10. In this section, we shall work out the details of the adjoint equation and the reduced gradient for optimization problems constrained by partial differential equations (PDE-constrained optimization). For high-order optimization methods, such as Newton-type techniques with Krylov subspace approaches, we shall apply the development in Chapter 20 to derive exact Hessian-vector products using adjoint. We consider prototype steady state (time-independent) PDEs of elliptic and hyperbolic types. The goal is to show how we translate abstract results in Lemma 9.4 and section 20.2 to concrete problems. This can serve as the baseline for carrying out the same task for different PDE-constrained optimization problems. Other topics on PDE-constrained optimization can be found in [96, 22, 21, 23, 92].

17.1 An advection-PDE-constrained optimization problem

Consider the following PDE-constrained optimization problem

$$\min_{z,u} J := \|u\|^2_{\mathbb{L}^2(\Omega)} := \frac{1}{2}\int_\Omega u^2 \, d\Omega$$

subject to

$$\begin{aligned} \boldsymbol{\beta}\cdot\nabla u &= 0, \quad \text{in } \Omega, \\ \boldsymbol{\beta}\cdot\boldsymbol{n} u &= z, \quad \text{in } \partial\Omega_{\text{in}}, \end{aligned} \tag{17.1}$$

where $u \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) := \left\{u : u \in \mathbb{L}^2(\Omega) \text{ and } \boldsymbol{\beta}\cdot\nabla u \in \mathbb{L}^2(\Omega)\right\}$. See Example 12.9 for the definition of other quantities in the constraint and the associated differential operator together with its adjoint. This optimization problem is a special case of the abstract problem discussed in Lemma 9.4. Note that for $u \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)$, its trace (in fact weighted trace with weight $|\boldsymbol{\beta}\cdot\boldsymbol{n}|$) on $\partial\Omega_{\text{in}}$ belongs to $\mathbb{L}^2(\partial\Omega_{\text{in}})$. The correct space for z is thus $\mathbb{L}^2(\partial\Omega_{\text{in}})$ with a weighted norm (see section 17.1). Since the constraints map $[u,z] \in \mathbb{X}\times\mathbb{Z} := \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)\times\mathbb{L}^2(\partial\Omega_{\text{in}})$ to $\mathbb{Y}:=\mathbb{L}^2(\Omega)\times\mathbb{L}^2(\partial\Omega_{\text{in}})$, the Lagrange multiplier $y = [v,w]$ has two components $v \in \mathbb{L}^2(\Omega)$ and $w \in \mathbb{L}^2(\partial\Omega_{\text{in}})$, respectively. Our task is to find the explicit form of the first order optimality condition (9.10) which, we recall, is a special case of the first order optimality condition via the Lagrangian multiplier Theorem 9.3. For practical PDE-constrained problem, the adjoint operators $[\mathscr{D}_u c(u_0,z_0)]^* y$ and $[\mathscr{D}_z c(u_0,z_0)]^* y$ are subtlely coupled and we have to go back to the Lagrangian functional in Theorem 9.3 to derive the optimality condition. To that end, let us form the Lagrangian functional

$$L(z,u) = \frac{1}{2}\int_\Omega u^2\, d\Omega + \int_\Omega (\boldsymbol{\beta}\cdot\nabla u)\, v\, d\Omega + \int_{\partial\Omega_{\text{in}}} (\boldsymbol{\beta}\cdot\boldsymbol{n}u - z)\; w\, ds,$$

and note that our optimization variable has two components $[u,z] \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)\times\mathbb{L}^2(\partial\Omega_{\text{in}})$. Take an arbitrary direction $[h,r] \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)\times\mathbb{L}^2(\partial\Omega_{\text{in}})$, the first order optimality condition (9.5), *with* $[u,z]$ *in place of* u_0 *and* $[h,r]$ *in place of* h, reads

$$\langle [v,w], \mathscr{D}c([u,z],[h,r])\rangle_{\mathbb{Y}} = \int_\Omega (\boldsymbol{\beta}\cdot\nabla h)\, v\, d\Omega + \int_{\partial\Omega_{\text{in}}} (\boldsymbol{\beta}\cdot\boldsymbol{n}h - r)\; w\, ds,$$

which after integration by parts becomes

$$\langle [v,w], \mathscr{D}c([u,z],[h,r])\rangle_{\mathbb{Y}} = -\int_\Omega (\boldsymbol{\beta}\cdot\nabla v)\, h\, d\Omega + \int_{\partial\Omega_{\text{in}}} \boldsymbol{\beta}\cdot\boldsymbol{n}\,(w+v)\, h\, ds + \int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta}\cdot\boldsymbol{n} v\, h\, ds - \int_{\partial\Omega_{\text{in}}} r\, w\, ds.$$

Here, we have restricted v in $\mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)$ for the differential and integral operators to make sense. The first order optimality condition (9.5) in this case reads: $\forall\, [h,r] \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega)\times\mathbb{L}^2(\partial\Omega_{\text{in}})$,

$$\int_\Omega uh\, d\Omega - \int_\Omega (\boldsymbol{\beta}\cdot\nabla v)\, h\, d\Omega + \int_{\partial\Omega_{\text{in}}} \boldsymbol{\beta}\cdot\boldsymbol{n}\,(w+v)\, h\, ds + \int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta}\cdot\boldsymbol{n} v\, h\, ds - \int_{\partial\Omega_{\text{in}}} r\, w\, ds = 0,$$

which, after taking $r = 0$ and any $h \in \mathbb{H}^1_{\boldsymbol{\beta},0}(\Omega) := \left\{u \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) : u|_{\partial\Omega} = 0\right\}$, becomes

$$\int_\Omega uh \, d\Omega - \int_\Omega (\boldsymbol{\beta} \cdot \nabla v) \, h \, d\Omega = 0, \quad \forall h \in \mathbb{H}^1_{\boldsymbol{\beta},0}(\Omega),$$

which implies[1]

$$-\boldsymbol{\beta} \cdot \nabla v + u = 0.$$

Consequently, the first order optimality condition reduces to: $\forall [h, r] \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) \times \mathbb{L}^2(\partial\Omega_{\text{in}})$,

$$\int_{\partial\Omega_{\text{in}}} \boldsymbol{\beta} \cdot n \, (w + v) \, h \, ds + \int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta} \cdot \boldsymbol{n} v \, h \, ds - \int_{\partial\Omega_{\text{in}}} r \, w \, ds = 0, \tag{17.2}$$

which, by taking $h = 0$ on $\partial\Omega_{\text{out}}$ and $r = 0$, becomes

$$\int_{\partial\Omega_{\text{in}}} \boldsymbol{\beta} \cdot \boldsymbol{n} \, (w + v) \, h \, ds = 0,$$

which in turn gives[2]

$$w = -v \text{ on } \partial\Omega_{\text{in}},$$

that is, the adjoint variables are not independent. This can be then substituted into (17.2) to further reduce the first order optimality condition to

$$\int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta} \cdot \boldsymbol{n} v \, h \, ds + \int_{\partial\Omega_{\text{in}}} r \, v \, ds = 0, \quad \forall [h, r] \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) \times \mathbb{L}^2(\partial\Omega_{\text{in}}).$$

It follows that, by taking $r = 0$ and using the surjectivity in section 17.1, we conclude

$$\boldsymbol{\beta} \cdot \boldsymbol{n} v = 0 \text{ on } \partial\Omega_{\text{out}},$$

and thus

$$v = 0 \text{ on } \partial\Omega_{\text{in}}.$$

In summary, the control equation (9.10c) becomes

$$v = 0 \text{ on } \partial\Omega_{\text{in}}, \tag{17.3}$$

and the adjoint equation (9.10b) reads

[1] It is due to the fact that $H^1_{\boldsymbol{\beta},0}(\Omega)$ is dense in $\mathbb{L}^2(\Omega)$ assuming Ω has segment property [8].

[2] Note that this is true due to the fact that the trace operator $\gamma : H^1_{\boldsymbol{\beta}}(\Omega) \to \mathbb{L}^2_{\boldsymbol{\beta}\cdot n}(\partial\Omega)$ is a continuous surjection (see, e.g., [31]), where $\mathbb{L}^2_{\boldsymbol{\beta}\cdot\boldsymbol{n}}(\partial\Omega)$ is $\mathbb{L}^2(\partial\Omega)$ with the weighted inner product $(u, v)_{\mathbb{L}^2_{\boldsymbol{\beta}\cdot\boldsymbol{n}}(\partial\Omega)} := \int_{\partial\Omega} |\boldsymbol{\beta} \cdot \boldsymbol{n}| \, uv \, ds$. When $|\boldsymbol{\beta} \cdot \boldsymbol{n}|$ is bounded on $\partial\Omega$, which is assumed for our setting, the two norms are equivalent, and thus it does not matter which norm we work with.

$$-\boldsymbol{\beta}\cdot\nabla v = -u \text{ in } \Omega, \tag{17.4a}$$
$$\boldsymbol{\beta}\cdot\boldsymbol{n}v = 0 \text{ on } \partial\Omega_{\text{out}}, \tag{17.4b}$$

Note that the differential operator on the left side of the adjoint equation (together with the homogeneous boundary condition) is exactly the adjoint operator we found in Example 12.9, which is not a surprise. As can be seen, the adjoint equation describes a reverse flow with $-\boldsymbol{\beta}$ velocity (compared to β in the forward equation) with (the derivative of) the objective function, particularly the forward solution u, as the forcing. The control equation says that at the optimal the forcing of the adjoint equation is such that the adjoint solution v on $\partial\Omega_{\text{in}}$ must vanish. Clearly, one admissible solution is that the adjoint is identically zero and the forcing u is identically zero. It then follows from the forward equation that $z = 0$. This is not surprising since, by inspection, the quadratic optimization under consideration has a solution $u = 0$ and $z = 0$ (different objective functions are presented in Problem 17.1. The reduced gradient can be now computed for a given z via three steps: 1) solve the *forward equation* (17.1) for $u(z)$, 2) solve the *adjoint equation* (17.4) for $v(u(z), z)$, and 3) substitute $v(u(z), z)$ into the left hand side of (17.3) to obtain the reduced gradient.

The next subject is how to compute the Hessian-vector products. The motivation for such a task in operator/matrix-free high-order optimization methods, e.g. Newton methods, can be referred to Chapter 20. To that end, we note that the reduced gradient

$$\nabla J = v \quad \text{ on } \partial\Omega_{\text{in}}$$

is the particular example of abstract reduced gradient (20.11). Similarly, the forward equation (17.1) the adjoint equation (17.4) are specific instances of (20.12a) and (20.12b), respectively: note that we have used v as the adjoint state instead of y. The question is how the Hessian-vector products in (20.13) and in (20.15) unfold for this particular optimization constrained by advection PDE (17.1). As shown above, the nontrivial part is the derivation of the reduced gradient (and thus the adjoint equation (17.4)) as they are not immediate from the abstract settings Lemma 9.4 and section 20.2. The Hessian-vector products, as disclosed in section 20.2, are then quite straightforward using directional derivatives.

Let $\hat{z}$ be an arbitrary function in $\mathbb{L}^2(\partial\Omega_{\text{in}})$. For notational convenience, let us denote the directional derivative of any quantity $(\cdot)$ with respect to z in the direction $\hat{z}$ as $\widehat{(\cdot)}$: for example, $\hat{z}$ is the directional derivative of z along $\hat{z}$ direction. Clearly, the key to realize is that the product of the (full) Hessian operator of J and $\hat{z}$, $\mathscr{H}_F\hat{z}$ is nothing more than the directional derivative of the gradient ∇J at z along the direction $\hat{z}$. Thus, we have

$$\mathscr{H}_F\hat{z} := \widehat{\nabla J} = \hat{v} \text{ on } \partial\Omega_{\text{in}}, \tag{17.5}$$

where $\hat{v}$, the directional derivative of v at z along the direction $\hat{z}$, can be obtained by differentiate the adjoint equation (17.4):

$$-\boldsymbol{\beta} \cdot \nabla \hat{v} = -\hat{u} \text{ in } \Omega, \tag{17.6a}$$

$$\boldsymbol{\beta} \cdot \boldsymbol{n} \hat{v} = 0 \text{ on } \partial\Omega_{\text{out}}, \tag{17.6b}$$

where $\hat{u}$, the directional derivative of u at z along the direction $\hat{z}$, can be obtained by differentiate the forward equation (17.1):

$$\begin{aligned} \boldsymbol{\beta} \cdot \nabla \hat{u} &= 0, \quad \text{in } \Omega, \\ \boldsymbol{\beta} \cdot \boldsymbol{n} \hat{u} &= \hat{z}, \quad \text{in } \partial\Omega_{\text{in}}. \end{aligned} \tag{17.7}$$

In summary, at the current[3] z and a given $\hat{z}$, we first solve (17.7) for $\hat{u}$, which is then substituted into (17.6) to solve for $\hat{v}$, and then finally compute the full Hessian-vector product (17.5). Since this problem is linear, and thus there are no second-order derivatives in the computation of the full Hessian, the Gauss-Newton Hessian is identical to the full Hessian. For nonlinear settings in which the Gauss-Newton Hessian is different from the full Hessian, we refer to Problem 17.2.

17.2 Elliptic-PDE-constrained optimization problem

In this section, we derive the gradient and Hessian-vector products for an optimization problem constrained by the elliptic PDE given in Example 12.7. In particular, consider

$$\min_{z,u} J(u) := \frac{1}{2} \int_{\Omega} \left(u - u^{obs}\right)^2 d\Omega$$

subject to

$$-\nabla \cdot (e^z \nabla u) = 0, \quad \text{in } \Omega, \tag{17.8a}$$

$$u = g, \quad \text{in } \partial\Omega, \tag{17.8b}$$

where $u^{obs}(\boldsymbol{x})$ is some reference/observational data, and the definition of the operator associated with the constraint and its adjoint are given in Example 12.7: in particular, $z \in \mathcal{C}^1(\Omega) \subset \mathbb{L}^2(\Omega)$. Here, g is the Dirichlet boundary data. This optimization problem is a special case of the abstract one in Lemma 9.4. Thus, the constraint maps $[u, z] \in \mathbb{X} \times \mathbb{Z} := \mathbb{H}_{\mathscr{A}} \times \mathcal{C}^1(\Omega)$ to $\mathbb{L}^2(\Omega) \times \mathbb{L}^2(\partial\Omega)$, and the Lagrange multiplier $y = [v, w]$ has two components $v \in \mathbb{L}^2(\Omega)$ and $w \in \mathbb{L}^2(\partial\Omega)$, respectively. Similar to section 17.1, we have to go back to Lagrangian multiplier Theorem 9.3 to derive the explicit form

[3] Here, we mean the value of z at the current optimization iteration.

of the first order optimality condition. In this case, the Lagrangian reads

$$L(z,u) = J(u) + \int_\Omega \left[-\nabla\cdot(e^z\nabla u)\right] v\, d\Omega + \int_{\partial\Omega} (u-g)\; w\, ds,$$

and note that our optimization variable has two components $[u,z] \in \mathbb{H}_{\mathscr{A}} \times \mathcal{C}^1(\Omega)$. Take an arbitrary direction $[h,r] \in \mathbb{H}_{\mathscr{A}} \times \mathcal{C}^1(\Omega)$, the first order optimality condition (9.5), *with* $[u,z]$ *in place of* u_0 *and* $[h,r]$ *in place of* h, reads

$$\int_\Omega \left(u-u^{obs}\right) h\, d\Omega + \int_\Omega \left[-\nabla\cdot(e^z\nabla h)\right] v\, d\Omega + \int_{\partial\Omega} h\, w\, ds + \int_\Omega \left[-\nabla\cdot(e^z r\nabla u)\right] v\, d\Omega = 0.$$

We next restrict $v \in \mathbb{H}_{\mathscr{A}}$ and integrate the second term by parts two times we arrive at:

$$\int_\Omega \left(u-u^{obs}\right) h\, d\Omega + \int_\Omega \left[-\nabla\cdot(e^z\nabla v)\right] h\, d\Omega + \int_{\partial\Omega} h\, w\, ds - \int_{\partial\Omega} e^z\nabla h\cdot\boldsymbol{n}\, v\, ds + \int_{\partial\Omega} e^z\nabla v\cdot\boldsymbol{n}\, h\, ds + \int_\Omega \left[-\nabla\cdot(e^z r\nabla u)\right] v\, d\Omega = 0, \quad \forall\, [h,r] \in \mathbb{H}_{\mathscr{A}} \times \mathcal{C}^1(\Omega). \tag{17.9}$$

Following a similar strategy[4] as in section 17.1 for (17.9) gives the adjoint equation

$$-\nabla\cdot(e^z\nabla v) = -\left(u-u^{obs}\right), \quad \text{in } \Omega, \tag{17.10a}$$

$$v = 0, \quad \text{in } \partial\Omega. \tag{17.10b}$$

Note that the differential operator on the left side of the adjoint equation (together with the homogeneous boundary condition) is exactly the adjoint operator we found in Example 12.7, which is not a surprise. The first order optimality condition is thus reduced to: $\forall\, [h,r] \in \mathbb{H}_{\mathscr{A}} \times \mathcal{C}^1(\Omega)$,

$$\int_{\partial\Omega} h\, w\, ds + \int_{\partial\Omega} e^z\nabla v\cdot\boldsymbol{n}\, h\, ds + \int_\Omega \left[-\nabla\cdot(e^z r\nabla u)\right] v\, d\Omega = 0, \tag{17.11}$$

which, by taking $r = 0$, gives,

$$w = -e^z\nabla v\cdot\boldsymbol{n}.$$

[4] Here we take $h \in \mathcal{C}_0^\infty(\Omega)$ and $r = 0$ to obtain the equation (17.10a) for v. To get the boundary condition (17.10b), we then take $h \in \mathcal{C}_0^1(\Omega)$ so that the normal trace $\nabla h\cdot\boldsymbol{n}$ is surjective on $\mathbb{L}^2(\partial\Omega)$.

Similar to section 17.1, we see that the second component of the adjoint variable y depends on the first, hence v is in fact the only adjoint variable. The first order optimality condition is further reduced to

$$\int_{\Omega} \left[-\nabla \cdot \left(e^z r \nabla u\right)\right] v \, d\Omega = 0, \quad \forall r \in \mathcal{C}^1\left(\Omega\right),$$

which—after integrating by parts, using the fact that $v = 0$ on $\partial\Omega$, and using the fact that $\mathcal{C}^1\left(\Omega\right)$ is dense in $L^2\left(\Omega\right)$—gives the control equation

$$e^z \nabla u \cdot \nabla v = 0. \tag{17.12}$$

The reduced gradient can be now computed for a given z via three steps: 1) solve the *forward equation* (17.8) for $u\left(z\right)$, 2) solve the *adjoint equation* (17.10) for $v\left(u(z), z\right)$, and 3) substitute $u(z)$ and $v(u\left(z\right), z)$ into the left hand side of (17.12) to obtain the reduced gradient

$$\nabla J = e^z \nabla u \cdot \nabla v.$$

The next task is to derive the Hessian-vector products for high-order optimization with Krylov subspace (operator/matrix free) methods. As in section 17.1, let $\hat{z} \in \mathcal{C}^1\left(\Omega\right) \subset \mathbb{L}^2\left(\Omega\right)$ and denote the directional derivative of any quantity $(\cdot)$ with respect to z in the direction $\hat{z}$ as $\widehat{(\cdot)}$.

Since the product of the (full) Hessian operator of J and $\hat{z}$, $\mathscr{H}_F \hat{z}$ is the directional derivative of the gradient ∇J at z along the direction $\hat{z}$, we have

$$\mathscr{H}_F \hat{z} := \widehat{\nabla J} = \underbrace{\hat{z} e^z \nabla u \cdot \nabla v + e^z \nabla \hat{u} \cdot \nabla v}_{\text{second order terms}} + e^z \nabla u \cdot \nabla \hat{v}, \tag{17.13}$$

where $\hat{u}$ and $\hat{v}$ are obtained from directional differentiation of the forward (17.8) and adjoint (17.10) equations. Doing so gives us the following equations to solve for $\hat{u}$ and $\hat{v}$:

$$\begin{aligned} -\nabla \cdot \left(e^z \hat{z} \nabla u\right) - \nabla \cdot \left(e^z \nabla \hat{u}\right) &= 0, \quad && \text{in } \Omega, \\ \hat{u} &= 0, \quad && \text{in } \partial\Omega, \end{aligned} \tag{17.14}$$

and

$$\begin{aligned} \underbrace{-\nabla \cdot \left(e^z \hat{z} \nabla v\right)}_{\text{second order term}} - \nabla \cdot \left(e^z \nabla \hat{v}\right) &= \hat{u}, \quad && \text{in } \Omega, \\ \hat{v} &= 0, \quad && \text{in } \partial\Omega. \end{aligned} \tag{17.15}$$

As shown in Chapter 20, the Gauss-Newton Hessian-vector product can be obtained by removing second-order derivative terms from the full Hessian-vector product (17.13) and the associated system (17.14) and (17.15).

In summary, at the current z and a given $\hat{z}$, we first solve (17.14) for $\hat{u}$, which is then substituted into (17.15) to solve for $\hat{v}$, and then finally compute

the full Hessian-vector product (17.13). The Gauss-Newton Hessian-vector product follows the same steps with the second-order terms removed.

Problems

Problem 17.1. Derive the reduced gradient and Hessian-vector products for the advection-PDE-constrained problem in section 17.1 with the following objective functions:

1. Let $u^{obs} \in \mathbb{L}^2(\Omega)$ be given and define
$$J := \frac{1}{2}\int_{\Omega} \left(u - u^{obs}\right)^2 d\Omega.$$
2. Let $\varphi \in \mathbb{L}^2(\Omega)$ and $f^{obs} \in$ be given and define
$$J := \left|(\varphi, u)_{\mathbb{L}^2(\Omega)}\right|^2$$

Problem 17.2. Consider two objective functions in Problem 17.1. For each case, derive the reduced gradient and Hessian-vector products for the advection-PDE-constrained problem in section 17.1 when z is given, but the optimization variable is $\boldsymbol{\beta}$.

Problem 17.3. Derive the reduced gradient and Hessian-vector products for the elliptic-PDE-constrained problem in section 17.2 with the following objective functions:

1. Let $\mathcal{D}$ be a open subset of Ω and define
$$J(u) := \frac{1}{2}\int_{\mathcal{D}} \left(u - u^{obs}\right)^2 d\Omega,$$
2. Let $\left\{u_j^{obs}\right\}_{j=1}^N$ be the observational data at the points $\boldsymbol{x}_j, j = 1, \ldots, N$, and define
$$J := \frac{1}{2}\sum_{j=1}^{N} \left(u(\boldsymbol{x}_j) - u_j^{obs}\right)^2.$$

Chapter 18
Efficient gradient computation for Neural ordinary differential equations with adjoint

Abstract

Part IV
Additional Topics

Use the template *part.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your part title page and, if desired, a short introductory text (maximum one page) on its verso page in the Springer layout.

Chapter 19
The development of kernel methods in Support Vector Machines

Abstract
Look at my lecture notes.

1. Start with regression problem and derive the adjoint problem and then notice the kernel, just like we did in the lecture notes
2. then move the SVM formulation and then its adjoint. Need to introduce/prove the weak duality. Do we need to prove strong duality? then show that the SVM similarly only depends on the inner product of the data. Then say that we shall limit to the classification problem.
3. present a couple of problems (just like I did on the white board in class for 1D and 2D) to demonstrate that going to high dimension things become separable. In this case, we can guess the feature map, and hence the inner product in featured space, and hence the definition of the kernel
4. Now turn around and start with kernel first without knowing the feature map, and then show how to do SVM in the feature space without going there as we did in the class.

Chapter 20
Exact computation of Hessian-vector product

Abstract In Chapter 9 we presented an optimization theory for sufficiently smooth problems possessing up to second-order Fréchet derivatives of the cost functional. We derived an efficient adjoint approach to compute the reduced gradient in Lemma 9.4 that can be used in a gradient-based approach to solve equality-constrained optimization problems. This chapter presents an adjoint approach to exactly compute the Hessian-vector product of the cost functional. This is required for Newton-type approaches, such as the Newton conjugate gradient approach. The additional expense on Hessian-vector products is justified due to the quadratic convergence of the Newton approach close to a minimum. We shall show that computing a Hessian-vector product is not more expensive than computing the gradient. It is in fact cheaper if the forward equation (9.12a) is nonlinear as we need to solve linearized versions of (9.12a) and (9.12b). We shall also show how to derive the product of Gauss-Newton Hessian with an arbitrary vector exactly using adjoint approach. This is useful for many optimization methods that rely on the Gauss-Newton Hessian, a simplification of the full Hessian, such as the Gauss-Newton method. The chapter starts with a finite-dimensional setting, deriving the full and Gauss-Newton Hessians, and extends the results to an abstract setting with examples. The Hessian-vector products for training deep neural network and for optimization problems with PDE constraints will be presented in Chapter 10, and Chapter 17, respectively.

20.1 Finite dimensional setting

We begin with the following unconstrained least square optimization problem

$$\min_{\boldsymbol{z}\in\mathbb{R}^p} J(\boldsymbol{z}) := \frac{1}{2}\left\|\boldsymbol{f}^{obs} - \boldsymbol{f}(\boldsymbol{z})\right\|^2_{\mathbb{R}^m}, \tag{20.1}$$

where $\boldsymbol{f}^{obs} \in \mathbb{R}^m$ is given and $\boldsymbol{f} : \mathbb{R}^p \ni \boldsymbol{z} \mapsto \boldsymbol{f}(\boldsymbol{z}) \in \mathbb{R}^m$. To solve (20.1), we consider Newton-type approaches of the following form:

$$\boldsymbol{z}^{k+1} = \boldsymbol{z}^k - \boldsymbol{H}^{-1}\left(\boldsymbol{z}^k\right) \nabla_{\boldsymbol{z}} J\left(\alpha^k\right), \quad k = 1, 2, \ldots, \tag{20.2}$$

where $\boldsymbol{H}(\boldsymbol{z}) \in \mathbb{R}^{n \times n}$ is either the Hessian $\nabla_{\boldsymbol{z}}^2 J(\boldsymbol{z})$ or an approximation to the Hessian. If we define $\Delta \boldsymbol{z}^k = \boldsymbol{z}^{k+1} - \boldsymbol{z}^k$, the iteration in (20.2) is equivalent to

$$\boldsymbol{H}\left(\boldsymbol{z}^k\right) \Delta \boldsymbol{z}^k = -\nabla_{\boldsymbol{z}} J\left(\boldsymbol{z}^k\right), \quad \boldsymbol{z}^{k+1} = \boldsymbol{z}^k + \Delta \boldsymbol{z}^k. \tag{20.3}$$

The crux of Newton-type methods is to solve the linear system of equations (20.3) for $\Delta \boldsymbol{z}^k$. For large-scale problems ($n \gg 1$) it is impractical to form the Hessian $\boldsymbol{H}\left(\boldsymbol{z}^k\right)$ at every Newton step, and thus methods (for example Gaussian elimination) that need explicit construction of the Hessian is not feasible. Fortunately, Krylov-type approaches, such as the conjugate gradient method, require only the product of the Hessian with vectors. As long as we can compute the Hessian-vector products in a matrix-free fashion, these methods are readily applicable. In this chapter, the Hessian $\boldsymbol{H}(\boldsymbol{z})$ in (20.3) is either the full Hessian $\boldsymbol{H}_F(\boldsymbol{z})$ or its Gauss-Newton approximation $\boldsymbol{H}_{GN}(\boldsymbol{z})$ (called Gauss-Newton Hessian) and our objective is to develop adjoint methods to compute Hessian-vector products exactly without forming the Hessian. Without loss of generality, we only need to compute the gradient and Hessian-vector products at a given $\boldsymbol{z}$. When it is clear in the context, we also omit the dependence on $\boldsymbol{z}$ for simplicity in writing. To begin, define $\boldsymbol{r}(\boldsymbol{z}) := \boldsymbol{f}^{obs} - \boldsymbol{f}(\boldsymbol{z})$. By the chain rule, the jth component of $\nabla_{\boldsymbol{z}} J$ is given as

$$\frac{\partial J}{\partial \boldsymbol{z}_j} = \sum_{i=1}^m \boldsymbol{r}_i \frac{\partial \boldsymbol{r}_i}{\partial \boldsymbol{z}_j} = -\sum_{i=1}^m \boldsymbol{r}_i \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{z}_j},$$

or equivalently as

$$\nabla_{\boldsymbol{z}} J = \left(\nabla_{\boldsymbol{z}} \boldsymbol{f}\right)^T \boldsymbol{r}. \tag{20.4}$$

Recall our conventions in Chapter 9 that all vectors are column vectors, the gradient of scalar-valued function, e.g. $\nabla_{\boldsymbol{z}} J$, is a column vector, and the Jacobian of a vector-valued function, e.g. $\nabla_{\boldsymbol{z}} \boldsymbol{f}$, is a matrix whose ith row is the transpose of the gradient of the ith component of the function.

Next, we derive the full Hessian and the Gauss-Newton Hessian to show their difference. The $j\ell$ component of the full Hessian is, again by the chain rule, given as

$$\boldsymbol{H}_F(j, \ell) = \sum_{i=1}^m \left(\frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{z}_\ell} \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{z}_j} - \boldsymbol{r}_i \frac{\partial^2 \boldsymbol{f}_i}{\partial \boldsymbol{z}_j \partial \boldsymbol{z}_\ell} \right).$$

The $j\ell$ component of the Gauss-Newton Hessian is given by

$$\boldsymbol{H}_{GN}(j,\ell) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{z}_\ell} \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{z}_j},$$

or equivalently

$$\boldsymbol{H}_{GN} = (\nabla_{\boldsymbol{z}} \boldsymbol{f})^T \nabla_{\boldsymbol{z}} \boldsymbol{f},$$

which is a simplification of the full Hessian in which we ignore the second-order derivative terms. Clearly, the Gauss-Newton Hessian is the full Hessian if $\boldsymbol{f}$, and hence $\boldsymbol{r}$, is linear in $\boldsymbol{z}$. As a result, we can show that the Gauss-Newton Hessian is the full Hessian of an approximation of J when $\boldsymbol{r}$ is linearized at $\boldsymbol{z}$ (see Problem 20.1). It follows that the convergent rate of the Gauss-Newton method is at best second order.

Remark 20.1. We see that if the $\nabla_{\boldsymbol{z}} J$ can be expressed in the form (20.4), the Gauss-Newton Hessian is readily available.

We now extend unconstrained optimization to specific equality-constrained optimization problems similar to that in Example 9.8. Let us consider the following optimization problem

$$\min_{\boldsymbol{u}\in\mathbb{R}^n, \boldsymbol{z}\in\mathbb{R}^p} J(\boldsymbol{u},\boldsymbol{z}) := \frac{1}{2} \left\| \boldsymbol{f}^{obs} - \boldsymbol{f}(\boldsymbol{u},\boldsymbol{z}) \right\|_{\mathbb{R}^m}^2, \qquad \text{subject to } \mathbf{c}(\boldsymbol{u},\boldsymbol{z}) = \mathbf{0},$$

where $\boldsymbol{f} : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^m$ and $\mathbf{c}(\boldsymbol{u},\boldsymbol{z}) : \mathbb{R}^n \times \mathbb{R}^p \to \mathbb{R}^n$. We assume that $\det(\nabla_{\boldsymbol{u}} \mathbf{c}) \neq 0, \forall \boldsymbol{u}, \boldsymbol{z}$ so that the implicit function theorem Theorem 9.1 allows us to compute $\boldsymbol{u}$ as a function of $\boldsymbol{z}$ from the constraint. The optimization problem is thus unconstrained in the reduced variable $\boldsymbol{z}$ when $\boldsymbol{u}$ is implicitly expressed as a function of $\boldsymbol{z}$. We are interested in deriving the reduced gradient and the reduced Hessian-vector products of J at a given point $[\boldsymbol{u},\boldsymbol{z}]$. Applying Lemma 9.4 the (total) reduced gradient reads

$$\nabla J = -(\nabla_{\boldsymbol{z}} \boldsymbol{f})^T \boldsymbol{r} + (\nabla_{\boldsymbol{z}} \mathbf{c})^T \boldsymbol{v}, \tag{20.5}$$

where $\boldsymbol{r}(\boldsymbol{u},\boldsymbol{z}) := \boldsymbol{f}^{obs} - \boldsymbol{f}(\boldsymbol{u},\boldsymbol{z})$, and $\boldsymbol{u}$ and $\boldsymbol{v}$ are governed by the first order[1] forward and adjoint system:

$$\mathbf{c}(\boldsymbol{u},\boldsymbol{z}) = \mathbf{0}, \tag{20.6a}$$

$$-(\nabla_{\boldsymbol{u}} \boldsymbol{f})^T \boldsymbol{r} + (\nabla_{\boldsymbol{u}} \mathbf{c})^T \boldsymbol{v} = \mathbf{0}, \tag{20.6b}$$

For computing the Hessian-vector products it is important to recall Remark 9.3. Given $\boldsymbol{z}$ (e.g. at the current optimization step in the reduced space), (20.6a) allows us to compute $u(\boldsymbol{z})$ as a function of $\boldsymbol{z}$. We can then compute $\boldsymbol{v}(\boldsymbol{z})$ as a function of $\boldsymbol{z}$ from (20.6b). The gradient of J with respect to $\boldsymbol{z}$ is then computed using (20.5). In other words, (20.6) is simply a means to

[1] We call (20.6) as the "first order" system as it is resulted from the first derivatives/variation of a Lagrangian (see Corollary 9.2).

express $\boldsymbol{u}$ and $\boldsymbol{v}$ as a function of a given $\boldsymbol{z}$, so that J can be considered as a function of only the reduced variable $\boldsymbol{z}$.

We next derive the (reduced) Hessian-vector products. To that end, let $\hat{\boldsymbol{z}} \in \mathbb{R}^p$ be arbitrary and denote the total directional derivative of any quantity $(\cdot)$ with respect to $\boldsymbol{z}$ in the direction $\hat{\boldsymbol{z}}$ as $\widehat{(\cdot)}$: for example, $\hat{\boldsymbol{u}}$ is the directional derivative of $\boldsymbol{u}$ with respect to $\boldsymbol{z}$ in the direction $\hat{\boldsymbol{z}}$. Clearly, the product of the (full) Hessian of J and $\hat{\boldsymbol{z}}$, $\boldsymbol{H}_F\hat{\boldsymbol{z}}$ is nothing more than the directional derivative of the gradient ∇J in (20.5) along the direction $\hat{\boldsymbol{z}}$, i.e.,

$$\boldsymbol{H}_F\hat{\boldsymbol{z}} = \widehat{\nabla J} = -\widehat{\nabla_{\boldsymbol{z}}\boldsymbol{f}}^T\boldsymbol{r} + (\nabla_{\boldsymbol{z}}\boldsymbol{f})^T\,\widehat{\boldsymbol{f}} + \widehat{\nabla_{\boldsymbol{z}}\mathbf{c}}^T\boldsymbol{v} + (\nabla_{\boldsymbol{z}}\mathbf{c})^T\,\hat{\boldsymbol{v}}, \tag{20.7}$$

where, by the chain rule,

$$\begin{aligned}
\widehat{\nabla_{\boldsymbol{z}}\boldsymbol{f}} &= \nabla_{\boldsymbol{z}}\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\right)\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\right)\hat{\boldsymbol{u}},\\
\widehat{\boldsymbol{f}} &= \nabla_{\boldsymbol{z}}\boldsymbol{f}\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\boldsymbol{f}\hat{\boldsymbol{u}},\\
\widehat{\nabla_{\boldsymbol{z}}\mathbf{c}} &= \nabla_{\boldsymbol{z}}\left(\nabla_{\boldsymbol{z}}\mathbf{c}\right)\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{z}}\mathbf{c}\right)\hat{\boldsymbol{u}}.
\end{aligned}$$

Here, recall the Fréchet gradient in Definition 9.3, we have identified, for example, that

$$\begin{aligned}
\nabla_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\right)\hat{\boldsymbol{u}} &= \left\langle \mathscr{D}_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\right), \hat{\boldsymbol{u}}\right\rangle_{\mathbb{R}^n},\\
\nabla_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{z}}\mathbf{c}\right)\hat{\boldsymbol{u}} &= \left\langle \mathscr{D}_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{z}}\mathbf{c}\right), \hat{\boldsymbol{u}}\right\rangle_{\mathbb{R}^n}
\end{aligned}$$

are the directional derivative of $\nabla_{\boldsymbol{z}}\boldsymbol{f}$ and $\nabla_{\boldsymbol{z}}\mathbf{c}$ with respect to $\boldsymbol{u}$ in the direction of $\hat{\boldsymbol{u}}$. Since we are given $\boldsymbol{f}$ and $\boldsymbol{c}$, all the quantities involved in $\boldsymbol{H}_F\hat{\boldsymbol{z}}$ are known except $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$. As discussed above, they are a function of $\boldsymbol{z}$ via (20.6) and thus they can be computed by taking the directional derivative along the direction $\hat{\boldsymbol{z}}$ for both sides of (20.6):

$$\nabla_{\boldsymbol{u}}\mathbf{c}\,\hat{\boldsymbol{u}} + \nabla_{\boldsymbol{z}}\boldsymbol{c}\,\hat{\boldsymbol{z}} = \mathbf{0} \tag{20.8a}$$

$$\begin{aligned}
-\left(\nabla_{\boldsymbol{z}}\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\right)\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\right)\hat{\boldsymbol{u}}\right)^T\boldsymbol{r} + \left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\boldsymbol{f}\hat{\boldsymbol{u}}\right)\\
+\left(\nabla_{\boldsymbol{z}}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\left(\nabla_{\boldsymbol{u}}\boldsymbol{c}\right)\hat{\boldsymbol{u}}\right)^T\boldsymbol{v} + \left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^T\hat{\boldsymbol{v}} = \mathbf{0}.
\end{aligned} \tag{20.8b}$$

Note that the *second order system* (20.8), from which we can compute $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ as a function of $\hat{\boldsymbol{z}}$, is a linearization of the first order system (20.6).

The product of the Gauss-Newton Hessian with $\hat{\boldsymbol{z}}$ can be derived in two ways. The first approach is based on the observation that the Gauss-Newton Hessian is obtained by removing all second-order derivative terms in the full Hessian. In particular, from (20.7) we have

$$\boldsymbol{H}_{GN}\hat{\boldsymbol{z}} = \left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\boldsymbol{f}\hat{\boldsymbol{u}}\right) + \left(\nabla_{\boldsymbol{z}}\mathbf{c}\right)^T\hat{\boldsymbol{v}}, \tag{20.9}$$

where $\hat{\boldsymbol{u}}$ and $\hat{\boldsymbol{v}}$ can be computed from the second order (20.10) by removing all second derivative terms, i.e,

$$\nabla_{\boldsymbol{u}}\mathbf{c}\,\hat{\boldsymbol{u}} + \nabla_{\boldsymbol{z}}\mathbf{c}\hat{\boldsymbol{z}} = \mathbf{0} \tag{20.10a}$$

$$\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\hat{\boldsymbol{z}} + \nabla_{\boldsymbol{u}}\boldsymbol{f}\hat{\boldsymbol{u}}\right) + \left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^T\hat{\boldsymbol{v}} = \mathbf{0}. \tag{20.10b}$$

The second approach is based on Remark 20.1.

Proposition 20.1. *Let $\boldsymbol{G}$ be the total derivative of $\boldsymbol{r}$ with respect to $\boldsymbol{z}$. The following hold:*

- $\boldsymbol{G} = \nabla_{\boldsymbol{u}}\boldsymbol{f}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\nabla_{\boldsymbol{z}}\mathbf{c} - \nabla_{\boldsymbol{z}}\boldsymbol{f}$,
- $\boldsymbol{H}_{GN}\hat{\boldsymbol{z}} = \boldsymbol{G}^T\boldsymbol{G}\hat{\boldsymbol{z}}$ *is the Hessian-vector product in* (20.9).

Proof. We begin by casting (20.5) into the form (20.4) and this can be achieved by eliminating $\boldsymbol{v}$ using the adjoint equation (20.6b):

$$\nabla J = -\left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\right)^T\boldsymbol{r} + \left(\nabla_{\boldsymbol{z}}\mathbf{c}\right)^T\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-T}\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\right)^T\boldsymbol{r} = \underbrace{\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\nabla_{\boldsymbol{z}}\mathbf{c} - \nabla_{\boldsymbol{z}}\boldsymbol{f}\right)^T}_{\boldsymbol{G}}\boldsymbol{r}.$$

from which the Gauss-Newton Hessian reads

$$\boldsymbol{H}_{GN} = \boldsymbol{G}^T\boldsymbol{G} = \left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\nabla_{\boldsymbol{z}}\mathbf{c} - \nabla_{\boldsymbol{z}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\nabla_{\boldsymbol{z}}\mathbf{c} - \nabla_{\boldsymbol{z}}\boldsymbol{f}\right).$$

Thus,

$$\begin{aligned}
\boldsymbol{H}_{GN}\hat{\boldsymbol{z}} &= \left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\nabla_{\boldsymbol{z}}\mathbf{c} - \nabla_{\boldsymbol{z}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\nabla_{\boldsymbol{z}}\mathbf{c} - \nabla_{\boldsymbol{z}}\boldsymbol{f}\right)\hat{\boldsymbol{z}} \\
&= \left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\nabla_{\boldsymbol{z}}\mathbf{c} - \nabla_{\boldsymbol{z}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\quad\underbrace{\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-1}\widehat{\boldsymbol{c}}}_{-\hat{\boldsymbol{u}}\text{ from (20.10a)}}\quad -\nabla_{\boldsymbol{z}}\boldsymbol{f}\hat{\boldsymbol{z}}\right), \\
&= \left(\nabla_{\boldsymbol{z}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\hat{\boldsymbol{u}} + \nabla_{\boldsymbol{z}}\boldsymbol{f}\hat{\boldsymbol{z}}\right) - \left(\nabla_{\boldsymbol{z}}\mathbf{c}\right)^T\underbrace{\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)^{-T}\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\right)^T\left(\nabla_{\boldsymbol{u}}\boldsymbol{f}\hat{\boldsymbol{u}} + \nabla_{\boldsymbol{z}}\boldsymbol{f}\hat{\boldsymbol{z}}\right)}_{-\hat{\boldsymbol{v}}\text{ from (20.10b)}},
\end{aligned}$$

which is exactly (20.9).

Example 20.1. We are going to derive the Hessian-vector products for the problem Example 9.9. The Hessian-vector product in (20.7) simplifies to

$$\boldsymbol{H}_F\hat{\boldsymbol{z}} = \boldsymbol{B}^T\hat{\boldsymbol{v}},$$

where $\hat{\boldsymbol{v}}$ can be computed from second order system (20.8), and in this case it reduces to

$$\boldsymbol{A}\hat{\boldsymbol{u}} + \boldsymbol{B}\hat{\boldsymbol{z}} = \boldsymbol{0},$$
$$\boldsymbol{C}^T\boldsymbol{C}\hat{\boldsymbol{u}} + \boldsymbol{A}^T\hat{\boldsymbol{v}} = \boldsymbol{0}.$$

Since both $\boldsymbol{f}$ and the constraint are linear in this example, the Gauss-Newton Hessian is exactly the Hessian.

20.2 General setting

We have constructed the Hessian-vector product for a general finite dimensional setting in section 20.1, which serves as the basis and the guideline for the following exposition in a general setting. As shall be seen, all the results are essentially the same except with Fréchet derivatives in place of the Fréchet gradients in Euclidean spaces. We shall skip the details and leave them as an exercise for the readers (see Problem 20.2). We begin with a general setting similar to that in Corollary 9.2:

$$\min_{u\in\mathbb{X}, z\in\mathbb{Z}} \left\| f^{obs} - f\left(u,z\right)\right\|_{\mathbb{W}}^2, \qquad \text{subject to } \; c\left(u,z\right) = 0,$$

where $c\left(\cdot,\cdot\right) : \mathbb{X}\times\mathbb{Z} \to \mathbb{V}$ and $f\left(\cdot,\cdot\right) : \mathbb{X}\times\mathbb{Z} \to \mathbb{W}$, with $\mathbb{X}, \mathbb{V}, \mathbb{Z}$ and $\mathbb{W}$ are Hilbert spaces. We assume that *the Fréchet derivative of the constraint with respect to u, i.e. $\mathscr{D}_u c\left(u,z\right) : \mathbb{X} \to \mathbb{V}$, is invertible* at any point $[u,z]$, so that, by the implicit function Theorem 9.1, we can express u as a function of z through the constraint. Our subject of interest is to derive the reduced gradient and the reduced Hessian-vector products of J at a given point $[u,z]$. Applying Lemma 9.4 the (total) reduced gradient reads

$$\nabla J = -\left(\mathscr{D}_z f\right)^* r + \left(\mathscr{D}_z c\right)^* v, \tag{20.11}$$

where $r\left(u,z\right) := f^{obs} - f\left(u,z\right)$, and u and v are governed by the first order system:

$$c\left(u,z\right) = 0, \tag{20.12a}$$
$$-\left(\mathscr{D}_u f\right)^* r + \left(\mathscr{D}_u c\right)^* v = 0, \tag{20.12b}$$

We now derive the product of the (reduced) Hessian with an arbitrary $\hat{z} \in \mathbb{Z}$. As above, let us denote the directional derivative of any quantity $(\cdot)$ with respect to z in the direction $\hat{z}$ as $\hat{(\cdot)}$: for example, $\hat{v}$ is the directional derivative of v with respect to z in the direction $\hat{z}$. Clearly, the product of the (full) Hessian operator of J and $\hat{z}$, $\mathscr{H}_F\hat{z}$ is nothing more than the directional derivative of the gradient ∇J in (20.11) along the direction $\hat{z}$, i.e.,

$$\mathscr{H}_F\hat{z} = -\left[\mathscr{D}_z\left(\mathscr{D}_z f\right)\hat{z} + \mathscr{D}_u\left(\mathscr{D}_z f\right)\hat{u}\right]^* r + \left(\mathscr{D}_z f\right)^*\left(\mathscr{D}_z f\hat{z} + \mathscr{D}_u f\hat{u}\right) + \left[\mathscr{D}_z\left(\mathscr{D}_z c\right)\hat{z} + \mathscr{D}_u\left(\mathscr{D}_z c\right)\hat{u}\right]^* v + \left(\mathscr{D}_z c\right)^*\hat{v}, \quad (20.13)$$

where $\hat{u}$ and $\hat{v}$ are the solution of the following second order system obtained from taking directional derivative of (20.12) with respect to z in the direction $\hat{z}$:

$$\mathscr{D}_u c\,\hat{u} + \mathscr{D}_z c\hat{z} = 0 \quad (20.14a)$$

$$-\left[\mathscr{D}_z\left(\mathscr{D}_u f\right)\hat{z} + \mathscr{D}_u\left(\mathscr{D}_u f\right)\hat{u}\right]^* r + \left(\mathscr{D}_u f\right)^*\left(\mathscr{D}_z f\hat{z} + \mathscr{D}_u f\hat{u}\right) + \left[\mathscr{D}_z\left(\mathscr{D}_u c\right)\hat{z} + \mathscr{D}_u\left(\mathscr{D}_u c\right)\hat{u}\right]^* v + \left(\mathscr{D}_u c\right)^*\hat{v} = 0. \quad (20.14b)$$

The Gaussian-Newton counterparts can be obtained by dropping second derivative terms in (20.13) and (20.14). Specifically, we have

$$\mathscr{H}_{GN}\hat{z} = \left(\mathscr{D}_z f\right)^*\left(\mathscr{D}_z f\hat{z} + \mathscr{D}_u f\hat{u}\right) + \left(\mathscr{D}_z c\right)^*\hat{v}, \quad (20.15)$$

where $\hat{u}$ and $\hat{v}$ satisfy the following simplified second order system:

$$\mathscr{D}_u c\,\hat{u} + \mathscr{D}_z c\hat{z} = 0 \quad (20.16a)$$

$$\left(\mathscr{D}_u f\right)^*\left(\mathscr{D}_z f\hat{z} + \mathscr{D}_u f\hat{u}\right) + \left(\mathscr{D}_u c\right)^*\hat{v} = 0. \quad (20.16b)$$

A general version of Proposition 20.1 can now be stated.

Proposition 20.2. *Let $\mathcal{G}$ be the total derivative of r with respect to z. The following hold:*

- $\mathcal{G} = \mathscr{D}_u f\left(\mathscr{D}_u c\right)^{-1}\mathscr{D}_z c - \mathscr{D}_z f$,
- $\mathscr{H}_{GN}\hat{z} = \mathcal{G}^T\mathcal{G}\hat{z}$ *is the Hessian-vector product in* (20.15).

Proof. See Problem 20.3.

Example 20.2. Recall Example 9.10 in which we provided the expression for the reduced gradient. We now apply the above results to write out the Hessian-vector products for this example. Note that $\boldsymbol{\alpha}$ plays the role of z (and thus $\mathbb{R}^n$ is in place of $\mathbb{Z}$). Keeping the similar convention of notations, for this example, the full Hessian-vector product in (20.13) reduces to

$$\mathscr{H}_F\widehat{\boldsymbol{\alpha}} = \int_\Omega \hat{v}\left(\boldsymbol{x}\right)\int_\Omega \nabla_{\boldsymbol{\alpha}} k\left(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\alpha}\right) u\left(\boldsymbol{y}\right)\, d\boldsymbol{y}\, d\boldsymbol{x} + \int_\Omega v\left(\boldsymbol{x}\right)\int_\Omega \nabla_{\boldsymbol{\alpha}} k\left(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\alpha}\right)\hat{u}\left(\boldsymbol{y}\right)\, d\boldsymbol{y}\, d\boldsymbol{x} + \int_\Omega v\left(\boldsymbol{x}\right)\int_\Omega \nabla_{\boldsymbol{\alpha}}\left(\nabla_\alpha k\left(\boldsymbol{x},\boldsymbol{y};\boldsymbol{\alpha}\right)\right)\widehat{\boldsymbol{\alpha}}\, u\left(\boldsymbol{y}\right)\, d\boldsymbol{y}\, d\boldsymbol{x},$$

and the system (20.14), for which $\hat{u}$ and $\hat{v}$ satisfy, specifically becomes

$$(\mathscr{I} + \mathsf{K})\,\hat{u} + \int_{\Omega} (\nabla_{\boldsymbol{\alpha}} k\,(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\alpha})\,, \widehat{\boldsymbol{\alpha}})_{\mathbb{R}^n}\, u\,(\boldsymbol{y})\; d\boldsymbol{y} = 0$$

$$(\varphi, \hat{u})_{\mathbb{L}^2(\Omega)}\, \varphi \;+\; (\mathscr{I} + \mathsf{K}^*)\,\hat{v} \;+\; \int_{\Omega} (\nabla_{\alpha} k\,(\boldsymbol{x}, \boldsymbol{y}; \alpha)\,, \widehat{\boldsymbol{\alpha}})_{\mathbb{R}^n}\, v\,(\boldsymbol{x})\; d\boldsymbol{x} \quad = \quad 0$$

Similarly, the Gauss-Newton Hessian-vector product (20.15), when applied to this example, becomes

$$\mathscr{H}_{GN} \widehat{\boldsymbol{\alpha}} = \int_{\Omega} \hat{v}\,(\boldsymbol{x}) \int_{\Omega} \nabla_{\boldsymbol{\alpha}} k\,(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\alpha})\, u\,(\boldsymbol{y})\; d\boldsymbol{y}\, d\boldsymbol{x},$$

and the system (20.16) reduces to

$$\begin{aligned} (\mathscr{I} + \mathsf{K})\,\hat{u} + \int_{\Omega} (\nabla_{\boldsymbol{\alpha}} k\,(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\alpha})\,, \widehat{\boldsymbol{\alpha}})_{\mathbb{R}^n}\, u\,(\boldsymbol{y})\; d\boldsymbol{y} &= 0, \\ (\varphi, \hat{u})_{\mathbb{L}^2(\Omega)}\, \varphi + (\mathscr{I} + \mathsf{K}^*)\,\hat{v} &= 0. \end{aligned}$$

Problems

Problem 20.1. Let $\boldsymbol{r}\,(\boldsymbol{z}) := \boldsymbol{f}^{obs} - \boldsymbol{f}\,(\boldsymbol{z})$ and its linearization at $\hat{\boldsymbol{z}}$ is given by $\boldsymbol{p}\,(\boldsymbol{\theta}) := \boldsymbol{f}^{obs} - \boldsymbol{f}\,(\hat{\boldsymbol{z}}) - \nabla_{\boldsymbol{z}}\, \boldsymbol{f}|_{\hat{\boldsymbol{z}}}\, \boldsymbol{\theta}$. Consider the following least square optimization problem

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^p} \hat{S}\,(\boldsymbol{\theta}) := \frac{1}{2}\, \|\boldsymbol{p}\,(\boldsymbol{\theta})\|^2_{\mathbb{R}^m}\,.$$

Show that the $j\ell$ component of the full Hessian is given by

$$\boldsymbol{H}_F\,(j, \ell) = \sum_{i=1}^{m} \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{z}_\ell}\bigg|_{\hat{\boldsymbol{z}}} \frac{\partial \boldsymbol{f}_i}{\partial \boldsymbol{z}_j}\bigg|_{\hat{\boldsymbol{z}}},$$

Or equivalently

$$\boldsymbol{H}_F = \nabla_{\boldsymbol{z}}\, \boldsymbol{f}|^T_{\hat{\boldsymbol{z}}}\, \nabla_{\boldsymbol{z}}\, \boldsymbol{f}|_{\hat{\boldsymbol{z}}}\,.$$

Problem 20.2. Using the Fréchet derivative and gradients in Chapter 9 to derive the results in section 20.2.

Problem 20.3. Prove Proposition 20.2

Problem 20.4. Derive the full and Gauss-Newton Hessian-vector products for Problem 9.4.

Chapter 21
Stability of ordinary differential equations via adjoint

Abstract

In this section, we provide a brief view on the role of adjoint in the study of stability of the equilibria of ordinary differential equations (ODEs). Most of our mathematical exposition follows [95], and we limit ourselves to autonomous systems of the form

$$\dot{\boldsymbol{x}} := \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}(\boldsymbol{x}), \tag{21.1}$$

where $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{f} : \mathsf{G} \subset \mathbb{R}^n \to \mathbb{R}^n$ is assumed to be continuous and locally Lipschitz. The domain G of $\boldsymbol{f}$ is assumed to be a nonempty open subset of $\mathbb{R}^n$. Due to the translational invariance of autonomous system, without loss of generality, we can assume that $\mathbf{0} \in \mathsf{G}$ is an equilibrium point, i.e. $\boldsymbol{f}(\mathbf{0}) = \mathbf{0}$. We use $\boldsymbol{x}_0$ to denote the initial condition and I to denote the maximal interval of existence for a solution of (21.1). All the norms $\|\cdot\|$ and inner products $(\cdot,\cdot)$ in this section are the standard Euclidean ones, and in this case Example 5.5 shows that the adjoint of a real matrix is simply its transpose. For matrices, the norm is the induced operator norm.

Definition 21.1 (Lyapunov stability). The equilibrium point $\mathbf{0}$ is stable (in the sense of Lyapunov) if for any $\varepsilon > 0$, $\exists \delta > 0$ such that for every (maximal) solution $\boldsymbol{x} : I \to \mathsf{G}$ such that $\boldsymbol{x}(0) \leq \delta$, we have $\boldsymbol{x}(t) \leq \varepsilon$ for all $t \in I \cap (0, \infty)$.

Theorem 21.1 (Lyapunov direct method). *If there exists an open neighborhood* U *of* $\mathbf{0}$ *and a continuous differentiable function* V *such that*

1. $V(\mathbf{0}) = 0$ *and* $V(\boldsymbol{z}) > 0$ *for all* $\boldsymbol{z} \in \mathsf{U} \setminus \{\mathbf{0}\}$, *and*
2. $V_{\boldsymbol{f}}(\boldsymbol{z}) := (\nabla V(\boldsymbol{z}), \boldsymbol{f}(\boldsymbol{z})) := \sum_{i=1}^{n} \frac{\partial V}{\partial \boldsymbol{z}_i} \boldsymbol{f}_i(\boldsymbol{z}) \leq 0$ *for all* $\boldsymbol{z} \in \mathsf{U}$.

Then $\mathbf{0}$ *is a stable equilibrium point of* (21.1).

Proof. See [95, Theorem 5.2].

Definition 21.2 (Asymptotic stability). The equilibrium $\mathbf{0}$ is attractive if there exists $\delta > 0$ such that for every $x_0 \in \mathsf{G}$ such that $\|\boldsymbol{x}_0\| \leq \delta$, then the solution $\boldsymbol{x}(t) \to \mathbf{0}$ as $t \to \infty$. We say $\mathbf{0}$ asymptotically stable (in the sense of Lyapunov) if it is both stable and attractive.

Theorem 21.2 (A sufficient condition for asymptotic stability). *Assume that there exists a neighborhood* U *of* $\boldsymbol{x}_0$ *and a continuously differentiable function* V *such that*

i $V(\mathbf{0}) = 0$ *and* $V(\boldsymbol{z}) > 0$ *for all* $\boldsymbol{z} \in \mathsf{U} \setminus \{\mathbf{0}\}$*, and* $V_{\boldsymbol{f}}(\boldsymbol{z}) \leq 0$ *for all* $\boldsymbol{z} \in \mathsf{U}$*, and*

ii $\mathbf{0}$ *is the inverse image of* $V_{\boldsymbol{f}}(\boldsymbol{z}) = 0$*, i.e.,* $V_{\boldsymbol{f}}^{-1}(0) = \mathbf{0}$*.*

Then $\mathbf{0}$ *is asymptotically stable.*

Proof. See [95, Theorem 5.15].

We next study the stability of systems of linear ODEs, and this is where the adjoint *comes into the picture.* We then infer the stability of nonlinear systems using the stability of their linearizations. To that end, we consider linear systems with $\mathsf{G} = \mathbb{R}^n$, $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, and

$$\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}. \tag{21.2}$$

Clearly, the solution of the (21.2) can be written as matrix exponential [66, 104, 73]

$$\boldsymbol{x}(t) = \exp(\boldsymbol{A}t)\,\boldsymbol{x}_0.$$

Definition 21.3 (Exponential stability). The equilibrium $\mathbf{0}$ is called exponentially stable if there exist $M \geq 1$ and $\alpha > 0$ such as

$$\|\exp(\boldsymbol{A}t)\,\boldsymbol{x}_0\| \leq M \exp(-\alpha t)\,\|\boldsymbol{x}_0\|, \quad \forall t \geq 0 \text{ and } \forall \boldsymbol{x}_0 \in \mathbb{R}^n.$$

Definition 21.4 (Hurwitz matrices). Let $\sigma(\boldsymbol{A})$ denote the spectrum (the collection of all eigenvalues) of $\boldsymbol{A}$. $\boldsymbol{A}$ is Hurwitz if $\sigma(\boldsymbol{A}) \subset \{\lambda \in \mathbb{C} : \Re(\lambda) < 0\}$.

Proposition 21.1. *Let* $\boldsymbol{A} \in \mathbb{R}^{n \times n}$*. The following statements are equivalent:*

i) $\boldsymbol{A}$ *is Hurwitz.*
ii) $\mathbf{0}$ *is an exponentially stable equilibrium of* (21.2)*.*
iii) $\mathbf{0}$ *is an asymptotically stable equilibrium of* (21.2)*.*

Proof. See [95, Proposition 5.25].

We are in the position to discuss one of the main results of this section.

Theorem 21.3 (Necessary and sufficient conditions for exponential stability). *$\boldsymbol{A} \in \mathbb{R}^{n\times n}$ is Hurwitz iff for each symmetric positive definite (SPD) matrix $\mathcal{Q} \in \mathbb{R}^{n\times n}$, the matrix equation*

$$\mathcal{P}\boldsymbol{A} + \boldsymbol{A}^*\mathcal{P} + \mathcal{Q} = 0$$

has a SPD solution $\mathcal{P} \in \mathbb{R}^{n\times n}$.

Proof. For the necessity, suppose $\boldsymbol{A}$ is Hurwitz. It follows from Proposition 21.1 that $\mathbf{0}$ is an exponentially stable equilibrium, that is, there exists $M > 0$ and $\alpha > 0$ such that

$$\|\exp(\boldsymbol{A}t)\| = \sup_{\boldsymbol{x}_0\in\mathbb{R}^n} \frac{\|\exp(\boldsymbol{A}t)\,\boldsymbol{x}_0\|}{\|\boldsymbol{x}_0\|} \le M\exp(-\alpha t), \quad t \ge 0.$$

Now for any SPD matrix $\mathcal{Q}$, let us define

$$\mathcal{P} := \int_0^\infty \exp(\boldsymbol{A}^*)\,\mathcal{Q}\exp(\boldsymbol{A}t)\;dt,$$

which is a well-defined matrix since

$$\|\mathcal{P}\| \le \int_0^\infty \|\exp(\boldsymbol{A}^*t)\|\,\|\mathcal{Q}\|\,\|\exp(\boldsymbol{A}t)\|\;dt \le M\,\|\mathcal{Q}\| \int_0^\infty \exp(-2\alpha t)\;dt < \infty,$$

where in the second inequality we have used the fact from Proposition 5.3 that the norm of a linear continuous operator is equal to the norm of its adjoint. $\mathcal{P}$ is SPD as

$$(\boldsymbol{x}, \mathcal{P}\boldsymbol{x}) = \int_0^\infty (\boldsymbol{x}, \exp(\boldsymbol{A}^*t)\,\mathcal{Q}\exp(\boldsymbol{A}t)\,\boldsymbol{x})\;dt = \int_0^\infty (\exp(\boldsymbol{A}t)\,\boldsymbol{x}, \mathcal{Q}\exp(\boldsymbol{A}t)\,\boldsymbol{x})\;dt,$$

and the fact that $\mathcal{Q}$ is SPD. Furthermore,

$$\mathcal{P}\boldsymbol{A} + \boldsymbol{A}^*\mathcal{P} = \int_0^\infty \frac{d\left(\exp(\boldsymbol{A}^*t)\,\mathcal{Q}\exp(\boldsymbol{A}t)\right)}{dt}\,dt = -\mathcal{Q},$$

where we have used the fact that $\exp(\boldsymbol{A}t)$, and hence $\exp(\boldsymbol{A}^*t)$, decays exponentially to 0.

For the sufficiency, Proposition 21.1 says that we only need to show that $\mathbf{0}$ is an asymptotically stable equilibrium. To that end, let us construct the following function

$$V(\boldsymbol{z}) := (\boldsymbol{z}, \mathcal{P}\boldsymbol{z}).$$

Clearly $V(\boldsymbol{z}) \ge 0$ for all $\boldsymbol{z} \in \mathbb{R}^n$, and $V(\boldsymbol{z}) = 0$ iff $\boldsymbol{z} = \mathbf{0}$ as $\mathcal{P}$ is SPD. Furthermore

$$V_{\boldsymbol{f}}(\boldsymbol{z}) = 2(\mathcal{P}\boldsymbol{z}, \boldsymbol{A}\boldsymbol{z}) = (\boldsymbol{z}, (\mathcal{P}\boldsymbol{A} + \boldsymbol{A}^*\mathcal{P})\,\boldsymbol{z}) = -(\boldsymbol{z}, \mathcal{Q}\boldsymbol{z}) \le 0.$$

Thus, by Theorem 21.2, $\mathbf{0}$ is asymptotically stable, and this ends the proof.

Let us now use Theorem 21.3 to study the stability of the equilibrium $\mathbf{0}$ of the general nonlinear system (21.1).

Hypothesis 21.1 (Nonlinearly purturbed linear ODE systems). *We assume that* $\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{h}(\boldsymbol{x})$ *where* $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ *and* $\boldsymbol{h} : \mathsf{G} \to \mathbb{R}^n$ *is continuous with*

$$\lim_{\boldsymbol{z}\to\mathbf{0}} \frac{\|\boldsymbol{h}(\boldsymbol{z})\|}{\|\boldsymbol{z}\|} = 0, \tag{21.3}$$

that is, $\boldsymbol{h}(\boldsymbol{z}) = o(\boldsymbol{z})$. *In other words,* $\boldsymbol{h}$ *approaches* $\mathbf{0}$ *faster than* $\boldsymbol{z}$.

Theorem 21.4 (Linear stability implies nonlinear stability). *Assume Hypothesis 21.1 holds. If* $\mathbf{0}$ *is an asymptotically stable equilibrium of (21.2), it is also an asymptotically stable equilibrium of (21.1).*

Proof. Suppose $\mathbf{0}$ is an asymptotically stable equilibrium of (21.2). Using Theorem 21.3 we can pick $\mathcal{Q} = I$, and form $V(\boldsymbol{z}) := (\boldsymbol{z}, \mathcal{P}\boldsymbol{z})$. It follows that $V(\mathbf{0}) = 0$ and $V(\boldsymbol{z}) > 0$ for all $\boldsymbol{z} \neq \mathbf{0}$ owing to the SPD property of $\mathcal{P}$. In order to use Theorem 21.2 to conclude the proof, we just need to show that $V_{\boldsymbol{f}}(\boldsymbol{z}) < 0$. We have

$$V_{\boldsymbol{f}}(\boldsymbol{z}) = 2(\mathcal{P}\boldsymbol{z}, \boldsymbol{A}\boldsymbol{z} + \boldsymbol{h}(\boldsymbol{z})) = -\|\boldsymbol{z}\|^2 + 2(\mathcal{P}\boldsymbol{z}, \boldsymbol{h}(\boldsymbol{z})) \leq -\|\boldsymbol{z}\|^2 + 2\|\mathcal{P}\|\,\|\boldsymbol{z}\|\,\|\boldsymbol{h}(\boldsymbol{z})\|,$$

where we have used $\mathcal{P}\boldsymbol{A} + \boldsymbol{A}^*\mathcal{P} = -I$ in the second equality. Next, using (21.3) we can pick a sufficient small neighborhood $\mathsf{U} = \boldsymbol{B}_\varepsilon(\mathbf{0})$ such that $\|\boldsymbol{h}(\boldsymbol{z})\| \leq \dfrac{\|\boldsymbol{z}\|}{4\|\mathcal{P}\|}$. Thus $V_{\boldsymbol{f}}(\boldsymbol{z}) < 0$ for all $\boldsymbol{z} \in \mathsf{U}$, and this ends the proof.

Theorem 21.4 reduces the stability analysis of an equilibrium of a nonlinear ODE system to an appropriate linear ODE counterpart, which is much simpler as linear algebra is then all we need for studying the stability. *We have also seen in Theorem 21.3 and Theorem 21.4 that adjoint plays the key role in establishing a necessary and sufficient condition for the stability of an equilibrium of an ODE system.*

Example 21.1 (Linearization of nonlinear ODE systems). We assume $\boldsymbol{f} : \mathsf{G} \to \mathbb{R}^n$ is differentiable at $\mathbf{0}$, and $\mathbf{0} \in \mathsf{G}$ is an equilibrium. Let us set

$$\boldsymbol{A} := \nabla \boldsymbol{f}(\mathbf{0}),$$

where the gradient is defined in Definition 9.3 and Example 9.2. The definition of Fréchet derivative in (9.2) implies (21.3) with $\boldsymbol{h}(\boldsymbol{x}) := \boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{A}\boldsymbol{x}$. Then by Theorem 21.4, the asymptotic stability of $\mathbf{0}$ for the linearized system implies the asymptotic stability of $\mathbf{0}$ for the original nonlinear system.

Example 21.2 (The asymptotic stability meaning of the basic reproduction in epidemic modeling). With the ever-increasing human population on every

part of the earth, the shrinkage in the natural habitat for plants and animals, and the shortage of natural resources such as water and food, the emergence of new and re-emergence of old infectious diseases are inevitable. Epidemic modeling plays a key role in forecasting how an infectious disease (such as SARS and COVID-19) spreads. This in turn facilitates informative decision-making to prevent a disease outbreak. The reproduction number $\mathfrak{R}_0$ provides epidemiologically meaningful criteria to predict an outbreak [116, 71, 25, 99, 47, 91, 41, 26, 72, 139]. Its popular definition is "the number of secondary cases one infected individual produces in a population consisting of only susceptibles". As a result, if $\mathfrak{R}_0 < 1$ the disease dies out but persists as an endemic (or goes on extinction) if $\mathfrak{R}_0 > 1$.

One of the most popular approaches to model disease dynamics is to use ODEs [25, 99, 47, 91, 41, 26, 72] in which, for example in an Susceptible-Exposed-Infectious-Recovered-Susceptible (SEIRS) model, the components of $\boldsymbol{x}$ in (21.1) are typically the fraction of susceptible, exposed, infected, and recovered within the population under consideration. The dying out of a disease corresponds to an outbreak returning to a disease-free state, while persistence to an endemic corresponds to a disease that remains in the population. Clearly, the disease-free state is an equilibrium of the ODE system if it is a meaningful representation of the disease dynamics. The dying out of a disease, therefore, corresponds to a perturbation from and then a return to disease-free equilibrium (DFE). This in turn should correspond to the asymptotic stability of the DFE. This is exactly a mathematical justification of the reproduction number. Figure 21.1 is our effort[1] in sketching the association, via the next generation matrix approach [76, 139, 45], of the reproduction number being less than unity and the asymptotic stability of $\mathbf{0}$ as a DFE of an abstract epidemic ODE model $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$. Note that the top arrow is the implication (due to Theorem 21.4) and the rest are equivalences. (See our work [6], and the references therein, for a detailed exposition of this correspondent for a new SEIRS epidemic model.) As can be seen, when the number of secondary cases one infected individual produces is less than one, the DFE is asymptotically stable and the solution of the epidemic ODE model approaches the DFE as time goes on. Epidemically speaking, the disease dies out.

[1] Unpublished notes.

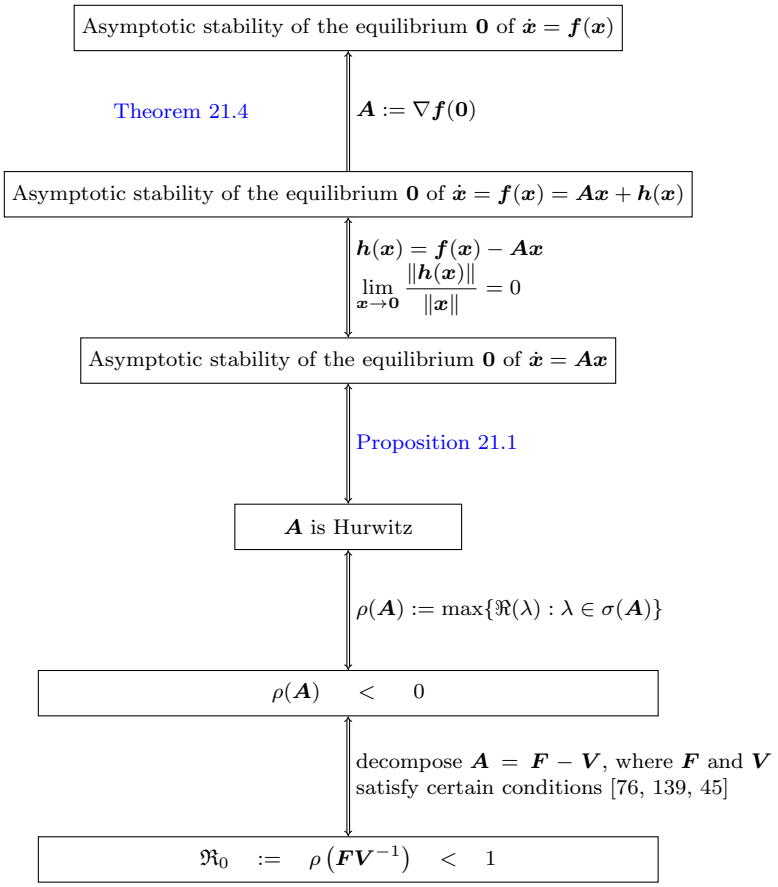


Fig. 21.1: A sketch of the association, via the next generation matrix approach, of the reproduction number being less than unity and the asymptotic stability of the disease-free equilibrium $\mathbf{0}$ of an abstract epidemic ODE model $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$. Note that the top arrow is the implication (due to Theorem 21.4) and the rest are equivalences.

Chapter 22
A new look at balanced truncation and its application to linear and nonlinear systems

Abstract In this chapter we provide a new look at the balanced truncation method [86, 105, 89, 9]. We explore the singular value decomposition. Though the starting point of our derivation is based on the principle component analysis as in [105], we provide several new perspectives. First, we view the reachability and observability from an operator point of view. Thus, we derive the reachability and observability gramians as, thanks to adjoint, a natural immediate step for computing the SVD of the input-to-state the state-to-output operators. Second, we use the calculus of variation (see **??**) to derive the reachability gramian in a natural way. Third, we provide an active subspace viewpoint of the most reachable and observable modes, and the balanced truncation.

22.1 Introduction

In this section, we discuss an important application of SVD in model order reduction (a.k.a. reduced-order modeling). For concreteness, we consider the following linear time-invariant systems (LTI)

$$\begin{cases} \dfrac{d\boldsymbol{x}}{dt} &= \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u} \\ \boldsymbol{y} &= \boldsymbol{C}\boldsymbol{x} \end{cases} \tag{22.1}$$

where $\boldsymbol{u} \in \mathbb{R}^p$ is some given input/control vector, $\boldsymbol{x} \in \mathbb{R}^n$ is the state, $\boldsymbol{y} \in \mathbb{R}^q$ is the output vector, $\boldsymbol{A} \in \mathbb{R}^{n\times n}$, $\boldsymbol{B} \in \mathbb{R}^{n\times p}$, and $\boldsymbol{C} \in \mathbb{R}^{q\times n}$ are given matrices, and t is time. The initial condition is given as $\boldsymbol{x}(t_0) = \boldsymbol{x}_0$ for some given $\boldsymbol{x}_0$. Note that considering a linear dynamical system in the matrix form (22.1) is not a limitation due to the equivalence of linear operators in finite dimensions and their matrix representations (see **??**). One of the most prominent model reduction methods of (22.1) is the so-called balanced truncation [86, 105, 89,

9]. We are interested in deriving the balanced truncation from SVD point of view. We begin with the analytical solution of (22.1) as

$$\boldsymbol{x}\left(t\right)=\underbrace{e^{\boldsymbol{A}\left(t-t_0\right)}\boldsymbol{x}_0}_{\boldsymbol{x}^I\left(t\right)}+\underbrace{\int_{t_0}^{t}e^{\boldsymbol{A}\left(t-\tau\right)}\boldsymbol{B}\boldsymbol{u}\left(\tau\right)\,d\tau}_{\boldsymbol{x}^{II}\left(t\right)}.$$

The first component $\boldsymbol{x}^I$ is the solution of (22.1) with zero control input $\boldsymbol{u}$. The corresponding output is

$$\boldsymbol{y}^I\left(t\right)=\boldsymbol{C}e^{\boldsymbol{A}\left(t-t_0\right)}\boldsymbol{x}_0.$$

The second component $\boldsymbol{x}^{II}$ is the solution is the solution of (22.1) with input $\boldsymbol{u}\left(t\right)$ and zero initial condition.

22.2 A constructive derivation of the observability Gramian

We fix $t_0=0$ and consider the output (vector-valued) function $\boldsymbol{y}^I\left(t\right)$, for $t\in\left(0,\infty\right)$, generated from x_0 with $\boldsymbol{u}=\theta$ (zero input). Let us define the following space square integrable vector-valued functions

$$\mathbb{L}^2\left(\mathbb{R}^q,\left(0,\infty\right)\right):=\left\{\boldsymbol{y}(t):\int_0^\infty\left\|\boldsymbol{y}\right\|_{\mathbb{R}^q}^2\,dt<\infty\right\}$$

with the inner product

$$\left(\boldsymbol{x},\boldsymbol{y}\right)_{\mathbb{L}^2\left(\mathbb{R}^q,\left(0,\infty\right)\right)}:=\int_0^\infty\left(\boldsymbol{x},\boldsymbol{y}\right)_{\mathbb{R}^n}\,dt,$$

and the induced norm

$$\left\|\boldsymbol{y}\right\|_{\mathbb{L}^2\left(\mathbb{R}^q,\left(0,\infty\right)\right)}=\sqrt{\left(\boldsymbol{y},\boldsymbol{y}\right)_{\mathbb{L}^2\left(\mathbb{R}^q,\left(0,\infty\right)\right)}}.$$

We assume that (22.1) is asymptotically stable, that is, $\boldsymbol{A}$ is Hurwitz (see Definition 21.2 and Definition 21.4). Define the map

$$\mathscr{O}:\mathbb{R}^n\ni\boldsymbol{x}_0\mapsto\mathscr{O}\boldsymbol{x}_0:=\boldsymbol{C}e^{\boldsymbol{A}t}\boldsymbol{x}_0\in\mathbb{L}^2\left(\mathbb{R}^q,\left(0,\infty\right)\right),$$

so that $\boldsymbol{y}^I\left(t\right)=\mathscr{O}\boldsymbol{x}_0$. Since $\boldsymbol{A}$ is Hurwitz, $\boldsymbol{y}^I\left(t\right)\in\mathbb{L}^2\left(\mathbb{R}^q,\left(0,\infty\right)\right)$. Thus, $\mathscr{O}$ maps an initial condition $\boldsymbol{x}_0$ to an output $\boldsymbol{y}^I$.

Definition 22.1 (Observability). Any $\boldsymbol{x}_0$ gives rise to $\boldsymbol{y}^I\left(t\right)\neq\theta$ is called observable. The system (22.1) is completely observable if every $\boldsymbol{x}_0\in\mathbb{R}^n$ is observable.

Since $\mathbb{R}^n$ is the domain of $\mathscr{O}$, its range is at most n-dimension, the operator $\mathscr{O}$ maps finite dimensional space $\mathbb{R}^n$ into a finite-dimensional subpsace of $\mathbb{L}^2(\mathbb{R}^q, (0,\infty))$. Clearly, $\boldsymbol{x}_0 \in \mathsf{N}(\mathscr{A}) \neq \{\theta\}$ is not observable as $\boldsymbol{y}^I(t) = \theta$. Thus, for all $\boldsymbol{x}_0 \in \mathbb{R}^n$ to be observable, $\mathsf{N}(\mathscr{A}) = \{\theta\}$. This is equivalent to the requirement that $\mathscr{O}$, as an $q \times n$ matrix for any t, has full column rank.

Thinking about the output $\boldsymbol{y}^I(t)$ as our gain from $\boldsymbol{x}_0$, we are interested in gaining the largest relative change in the norms (also called energy in this chapter) from $\boldsymbol{x}_0$ to $\boldsymbol{y}^I$. Since $\mathscr{O}$ is linear, this is the same as seeking $\boldsymbol{x}_0$ with unit energy that produces the largest output energy. Let us call that optimal initial condition as ϕ. As shown in section 8.2, ϕ must be the first principle component of $\mathscr{O}$ (or equivalently the first right singular vector), namely[1]

$$\phi = \arg \min_{\|\boldsymbol{x}_0\|_{\mathbb{R}^n}=1} (\mathscr{O}\boldsymbol{x}_0, \mathscr{O}\boldsymbol{x}_0)_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \arg \min_{\|\boldsymbol{x}_0\|_{\mathbb{R}^n}=1} \boldsymbol{x}_0^T \mathscr{O}^* \mathscr{O} \boldsymbol{x}_0,$$

that is, ϕ must be an eigenvector of $\mathscr{O}^*\mathscr{O}$ corresponding to the largest eigenvalue. Again, this is not a surprise as we have seen this in the proof of the SVD Theorem 8.1. We need to find $\mathscr{O}^*$ to see how $\mathscr{O}^*\mathscr{O}$ looks like. Using the definition of adjoint, for any $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{y}(t) \in \mathbb{L}^2(\mathbb{R}^q, (0,\infty))$ we have

$$(\mathscr{O}\boldsymbol{x}, \boldsymbol{y})_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \int_0^\infty \boldsymbol{y}(t)^T \boldsymbol{C} e^{\boldsymbol{A}t} \boldsymbol{x}\, dt = \left(\boldsymbol{x}, \int_0^\infty e^{\boldsymbol{A}^T t} \boldsymbol{C}^T \boldsymbol{y}(t)\, dt\right)_{\mathbb{R}^n},$$

which implies that $\mathscr{O}^* : \mathbb{L}^2(\mathbb{R}^q, (0,\infty)) \ni \boldsymbol{y}(t) \mapsto \int_0^\infty e^{\boldsymbol{A}^T t} \boldsymbol{C}^T \boldsymbol{y}(t)\, dt \in \mathbb{R}^n$. Thus, we have

$$\mathscr{O}^*\mathscr{O} = \int_0^\infty e^{\boldsymbol{A}^T t} \boldsymbol{C}^T \boldsymbol{C} e^{\boldsymbol{A}t}\, dt, \tag{22.2}$$

which is exactly the observability Gramian in the control literature (see, e.g., [9]). In other words, from the desire to find an initial condition $\boldsymbol{x}_0$ that yields the largest relative change in energy from $\boldsymbol{x}_0$ to $\boldsymbol{y}^I(t)$, we have constructively derived the observability Gramian using adjoint of the initial-condition-to-output map $\mathscr{O}$ and the SVD. The eigenvector ϕ corresponding to the largest eigenvalue of the observability Gramian is also known as the most observable initial condition. The corresponding largest output energy is thus given by

$$\left\|\boldsymbol{y}^I\right\|^2_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \phi^T \mathscr{O}^* \mathscr{O} \phi = \|\mathscr{O}\|^2.$$

[1] Again, as discussed in section 8.2, we should write $\phi \in \arg \min_{\|\boldsymbol{x}_0\|_{\mathbb{R}^n}=1} (\mathscr{O}\boldsymbol{x}_0, \mathscr{O}\boldsymbol{x}_0)_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))}$.

22.3 A constructive derivation of the reachability Gramian

We set $t_0 = 0$ and $t = \infty$ in $\boldsymbol{x}^{II}$, that is, we are interested in reaching $\boldsymbol{x}_0$ at time 0 by driving the LTI system (22.1) with $\boldsymbol{u}(t) \in \mathbb{L}^2(\mathbb{R}^p, (-\infty, 0))$ from negative infinity time with $\boldsymbol{x}(-\infty) = \boldsymbol{0}$. Here, the Hilbert space $\mathbb{L}^2(\mathbb{R}^p, (-\infty, 0))$ is defined similar to the definition of $\mathbb{L}^2(\mathbb{R}^q, (0, \infty))$ in section 22.2. In this case, $\boldsymbol{x}^{II}$ defines the following map

$$\mathscr{R} : \mathbb{L}^2(\mathbb{R}^p, (-\infty, 0)) \ni \boldsymbol{u}(t) \mapsto \boldsymbol{x}_0 = \mathscr{R}\boldsymbol{u} := \int_{-\infty}^{0} e^{-\boldsymbol{A}\tau} \boldsymbol{B}\boldsymbol{u}(\tau)\, d\tau \in \mathbb{R}^n.$$

Again, the definition of $\mathscr{R}$, and hence $\mathscr{R}$ is bounded, makes sense thanks to the fact that $\boldsymbol{A}$ is Hurwitz. A control question is then if there exists an input $\boldsymbol{u}(t)$ that drives the system (22.1) from zero state $\boldsymbol{x}(-\infty) = \boldsymbol{0}$ to the desired state $\boldsymbol{x}_0$. Such an $\boldsymbol{x}_0$ is call reachable.

Definition 22.2 (Reachability). We say that $\boldsymbol{x}_0$ is reachable if there exists $\boldsymbol{u} \in \mathbb{L}^2(\mathbb{R}^p, (-\infty, 0))$ such that $\boldsymbol{x}_0 = \mathscr{R}\boldsymbol{u}$. The system (22.1) is completely reachable if any $\boldsymbol{x}_0 \in \mathbb{R}^n$ is reachable.

Thus, reachability of $\boldsymbol{x}_0 \in \mathbb{R}^n$ is the same as existence of a solution $\boldsymbol{u} \in \mathbb{L}^2(\mathbb{R}^p, (-\infty, 0))$ of the equation

$$\mathscr{R}\boldsymbol{u} = \boldsymbol{x}_0,$$

and complete reachability is thus the subjectivity of $\mathscr{R}$, that is, $\mathsf{R}(\mathscr{R}) = \mathbb{R}^n$. From now on we assume that (22.1) is completely reachable. In this case, $\dim(\mathsf{R}(\mathscr{R}^*)) = \dim(\mathsf{R}(\mathscr{R})) = n$ (see **??**), and Corollary 5.2 says that $\mathsf{N}(\mathscr{R}) = \mathsf{R}(\mathscr{R}^*)^{\perp}$ must be infinite-dimensional subspace of $\mathbb{L}^2(\mathbb{R}^p, (-\infty, 0))$.

Now thinking $\boldsymbol{u}(t)$ as the cost (expense) of gaining $\boldsymbol{x}_0$ (gain), we aim to minimize the cost relative to gain. We use the norm of the input $\|\boldsymbol{u}\|_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}$ for cost and the norm $\|\boldsymbol{x}_0\|_{\mathbb{R}^n}$ for gain. We shall first the best input $\boldsymbol{u}(t)$ in two steps. First, we shall find an input with the smallest norm for each $\boldsymbol{x}_0$, and then identify which $\boldsymbol{x}_0$ produces the smallest input norm. For the first step, our task is

$$\min_{\boldsymbol{u}\in\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} \frac{\|\boldsymbol{u}\|_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}}{\|\boldsymbol{x}_0\|_{\mathbb{R}^n}}, \quad \text{subject to } \mathscr{R}\boldsymbol{u} = \boldsymbol{x}_0.$$

Since we are interested in the relative norm, we can simply consider $\boldsymbol{x}_0$ with unit-norm. Thus, our task of interest is an equality-constrained optimization problem

$$\min_{\boldsymbol{u}\in\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} \frac{1}{2}\|\boldsymbol{u}\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}, \quad \text{subject to } \mathscr{R}\boldsymbol{u} = \boldsymbol{x}_0. \tag{22.3}$$

Define the Lagrangian

$$L(\boldsymbol{u},\boldsymbol{z}) := \frac{1}{2}\|\boldsymbol{u}\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + (\boldsymbol{z},\mathscr{R}\boldsymbol{u}-\boldsymbol{x}_0)_{\mathbb{R}^n},$$

and apply the Lagrangian multiplier Theorem 9.3 we obtain

$$\boldsymbol{u}+\mathscr{R}^*\boldsymbol{z}=\theta,$$

which implies

$$\mathscr{R}\boldsymbol{u}+\mathscr{R}\mathscr{R}^*\boldsymbol{z}=\theta.$$

Now since $\dim(\mathsf{R}(\mathscr{R})) = n$ (complete reachability), $\mathscr{R}\mathscr{R}^* : \mathbb{R}^n \to \mathbb{R}^n$ is invertible (see Problem 22.1), and we can use the constraint to solve for the Lagrangian multiplier $\boldsymbol{z}$ as

$$\boldsymbol{z} = -(\mathscr{R}\mathscr{R}^*)^{-1}\boldsymbol{x}_0,$$

and thus the input with minimum norm that reaches a unit vector $\boldsymbol{x}_0$ is given by

$$\boldsymbol{u}(t) = \mathscr{R}^*(\mathscr{R}\mathscr{R}^*)^{-1}\boldsymbol{x}_0, \tag{22.4}$$

We are now ready for the second step in which we find an $\boldsymbol{x}_0$ whose associated $\boldsymbol{u}(t)$ given in (22.4) has the smallest norm. Since we have

$$\begin{aligned}\|\boldsymbol{u}(t)\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} &= \left(\mathscr{R}^*(\mathscr{R}\mathscr{R}^*)^{-1}\boldsymbol{x}_0, \mathscr{R}^*(\mathscr{R}\mathscr{R}^*)^{-1}\boldsymbol{x}_0\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}\\ &= \left(\boldsymbol{x}_0, (\mathscr{R}\mathscr{R}^*)^{-1}\mathscr{R}\mathscr{R}^*(\mathscr{R}\mathscr{R}^*)^{-1}\boldsymbol{x}_0\right)_{\mathbb{R}^n} = \boldsymbol{x}_0^T(\mathscr{R}\mathscr{R}^*)^{-1}\boldsymbol{x}_0\end{aligned}$$

A unit-norm $\boldsymbol{x}_0$ that we are looking for is thus a normalized eigenvector of $(\mathscr{R}\mathscr{R}^*)^{-1}$ associated with its smallest eigenvalue. In other words, a unit-norm $\boldsymbol{x}_0$ is a normalized eigenvector of $\mathscr{R}\mathscr{R}^*$ associated with the largest eigenvalue. An input with smallest energy is then given by

$$\|\boldsymbol{u}(t)\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} = \frac{1}{\text{the largest eigenvalue of } \mathscr{R}\mathscr{R}^*} = \frac{1}{\|\mathscr{R}\|^2}.$$

What remains is to determine $\mathscr{R}^*$ to know what the $n\times n$ matrix $\mathscr{R}\mathscr{R}^*$ looks like. Similar to section 22.2, using definition we find that (see Problem 8.6)

$$\mathscr{R}^* : \mathbb{R}^n \ni \boldsymbol{x}_0 \mapsto \boldsymbol{B}^T e^{-\boldsymbol{A}^T t}\boldsymbol{x}_0 \in \mathbb{L}^2(\mathbb{R}^p,(-\infty,0)),$$

and thus

$$\mathscr{R}\mathscr{R}^* = \int_{-\infty}^{0} e^{-\boldsymbol{A}\tau}\boldsymbol{B}\boldsymbol{B}^T e^{-\boldsymbol{A}^T\tau}\,d\tau = \int_0^{\infty} e^{\boldsymbol{A}\tau}\boldsymbol{B}\boldsymbol{B}^T e^{\boldsymbol{A}^T\tau}\,d\tau, \tag{22.5}$$

which is exactly the reachability Gramian in the control literature (see, e.g., [9]).

22.4 Balanced truncation

Recall from section 22.2 and section 22.3 that the output energy associated with a state $\boldsymbol{x}_0$ (with zero input $\boldsymbol{u}$)

$$\|\boldsymbol{y}\|^2_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \boldsymbol{x}_0^T \mathscr{O}^* \mathscr{O} \boldsymbol{x}_0,$$

and the (minimum) input energy that produces a state $\boldsymbol{x}_0$ (from zero state)

$$\|\boldsymbol{u}(t)\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} = \boldsymbol{x}_0^T \left(\mathscr{R}\mathscr{R}^*\right)^{-1} \boldsymbol{x}_0.$$

In this section, we assume the system (22.1) is both completely reachable and observable, and thus both $\mathscr{O}^*\mathscr{O}$ and $\mathscr{R}\mathscr{R}^*$ are invertible (see Problem 22.1). We also recall that the states $\boldsymbol{x}_0$ that are most observable are those align with eigenvectors of $\mathscr{O}^*\mathscr{O}$ (equivalently left right vectors of $\mathscr{O}$) corresponding to the largest eigenvalues (equivalently largest singular values of $\mathscr{O}$). On the other hand, the states $\boldsymbol{x}_0$ that are most reachable are those align with eigenvectors of $\mathscr{R}\mathscr{R}^*$ (equivalently left singular vectors of $\mathscr{R}$) corresponding to the largest eigenvalues (equivalently largest singular values of $\mathscr{R}$). Thus, in the eigenvector coordinates of $\mathscr{O}^*\mathscr{O}$ and $\mathscr{R}\mathscr{R}^*$ we know which states are important. For example, let the eigendecomposition of $\mathscr{O}^*\mathscr{O}$ be

$$\mathscr{O}^*\mathscr{O} = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T,$$

where Λ is a diagonal matrix whose diagonal entries are ordered from largest to smallest, and each column of $\boldsymbol{U}$ is the corresponding eigenvector. Moving to the eigenvector coordinates is equivalent to a linear transformation

$$\boldsymbol{x} = \boldsymbol{U}\hat{\boldsymbol{x}},$$

and the system (22.1) is transformed to

$$\begin{cases} \dfrac{d\hat{\boldsymbol{x}}}{dt} &= \hat{\boldsymbol{A}}\hat{\boldsymbol{x}} + \hat{\boldsymbol{B}}\boldsymbol{u}, \\ \boldsymbol{y} &= \hat{\boldsymbol{C}}\hat{\boldsymbol{x}}, \end{cases} \tag{22.6}$$

where $\hat{\boldsymbol{A}} = \boldsymbol{U}^T\boldsymbol{A}\boldsymbol{U}$, $\hat{\boldsymbol{B}} = \boldsymbol{U}^T\boldsymbol{B}$, and $\hat{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{U}$. The observability Gramian for (22.6) reads (see Problem 22.2)

$$\hat{\mathscr{O}}^*\hat{\mathscr{O}} = \boldsymbol{U}^T\mathscr{O}^*\mathscr{O}\boldsymbol{U} = \boldsymbol{\Lambda},$$

and the output energy now becomes

$$\|\boldsymbol{y}\|^2_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \hat{\boldsymbol{x}}_0^T \hat{\mathscr{O}}^* \hat{\mathscr{O}} \hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{x}}_0^T \boldsymbol{\Lambda} \hat{\boldsymbol{x}}_0 = \sum_{i=1}^{n} \lambda_i \hat{\boldsymbol{x}}_0^2(i)$$

which shows that the ith component of the transformed state $\hat{\boldsymbol{x}}_0$ contributes to the output energy more than, thus more observable than, the jth component. Let $\varepsilon > 0$ be small and there exists $n_r \in \mathbb{N}$ and $n_r < n$ such that $\lambda_i \leq \varepsilon$ for all $i \geq n_r$, then truncating the transformed system to have the following reduced system

$$\begin{cases} \dfrac{d\hat{\boldsymbol{x}}_r}{dt} &= \hat{\boldsymbol{A}}_r \hat{\boldsymbol{x}}_r + \hat{\boldsymbol{B}}_r \boldsymbol{u}, \\ \hat{\boldsymbol{y}}_r &= \hat{\boldsymbol{C}}_r \hat{\boldsymbol{x}}_r, \end{cases}$$

where $\hat{\boldsymbol{A}}_r$ is the $n_r \times n_r$ block of $\hat{\boldsymbol{A}}$ associated with $\hat{\boldsymbol{x}}_r$. Similarly, $\hat{\boldsymbol{B}}_r$ and $\hat{\boldsymbol{C}}_r$ are the subblocks of $\hat{\boldsymbol{B}}$ and $\hat{\boldsymbol{C}}$ associated with the reduced state $\hat{\boldsymbol{x}}_r$. Note that the output $\hat{\boldsymbol{y}}_r$ of the reduced system is not the same as the original output $\boldsymbol{y}$ due to truncation. However, the energy of the reduced output is Need to recheck the following identity: This is only true if the reduced gramians are the reduced diagonal of the original gramians

$$\|\boldsymbol{y}_r\|^2_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \sum_{i=1}^{n_r} \lambda_i \hat{\boldsymbol{x}}_0^2(i)\,.$$

Thus,

$$\|\boldsymbol{y}\|^2_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} - \|\boldsymbol{y}_r\|^2_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \sum_{i=n_r+1}^{n} \lambda_i \hat{\boldsymbol{x}}_0^2(i) \leq \varepsilon \sum_{i=n_r+1}^{n} \hat{\boldsymbol{x}}_0^2(i) \leq \varepsilon \|\boldsymbol{x}_0\|^2,$$

which means that the output energy loss due to truncation can be negligible if ε is small (i.e. when the eigenvalues of $\mathscr{O}^*\mathscr{O}$ decay quickly).

We can argue similarly, and thus ranking the new transformed states, by looking at the eigenvalue decomposition of $\mathscr{R}\mathscr{R}^*$. We can conclude that by removing the transformed states associated with small eigenvalues of $\mathscr{R}\mathscr{R}^*$ does not alter much the system behavior from input to state.

The question is then which coordinate transform we should use to rank the transform states, hence truncating the system appropriately? The problem here is that states the are more observable may be less reachable. Thus, keeping more observable states may not be meaningful as they could be hardly reachable. Conversely, keeping more reachable states may not be useful as they could be hardly observable. The intuitive approach is to transform the system so that any transformed state is equally observable and reachable. Then, by truncating less observable states, we also remove less reachable states at the same time. Such a transformation must simultaneously diagonalizes observability and reachability Gramians so that the transformed ob-

servability and reachability Gramians are not only diagonal but also equal to each other.

We can find a desirable transformed system in two steps:

1. First, transform the system (22.1) via $\boldsymbol{x} = \boldsymbol{S}\hat{\boldsymbol{x}}$, where $\boldsymbol{S} = \boldsymbol{U}\Lambda^{-1/2}$, then (see Problem 22.2) the transformed Gramians are

$$\hat{\mathscr{O}}^*\hat{\mathscr{O}} = I, \text{ and } \hat{\mathscr{R}}\hat{\mathscr{R}}^* = \Lambda^{1/2}\boldsymbol{U}^T\mathscr{R}\mathscr{R}^*\boldsymbol{U}\Lambda^{1/2}.$$

2. Second, let the eigenvalue decomposition of $\hat{\mathscr{R}}\hat{\mathscr{R}}^*$ as

$$\hat{\mathscr{R}}\hat{\mathscr{R}}^* = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T,$$

and transform the system once more time with $\hat{\boldsymbol{x}} = \boldsymbol{T}\tilde{\boldsymbol{x}}$ where $\boldsymbol{T} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$.

Then, both Gramians of the final transformed system are $\boldsymbol{\Sigma}$, which is exactly what we want. That is, any component of $\tilde{\boldsymbol{x}}$ is equally observable and reachable. Furthermore $\tilde{\boldsymbol{x}}(i)$ is more observable and more reachable than $\tilde{\boldsymbol{x}}(j)$ for $i > j$, where $i = 1, \ldots, n$ and $j = 1, \ldots, n$.

We point out that without truncation the output of the system remains unchanged through any linear and invertible transformation $\boldsymbol{S}$. It is also easy to see that (see Problem 22.2) the original cross Gramian $\mathscr{R}\mathscr{O} : \mathbb{R}^n \to \mathbb{R}^n$ and the transformed one $\hat{\mathscr{R}}\hat{\mathscr{O}} : \mathbb{R}^n \to \mathbb{R}^n$ are related by a similarity transformation

$$\hat{\mathscr{R}}\hat{\mathscr{O}} = \boldsymbol{S}^{-1}\mathscr{R}\mathscr{O}\boldsymbol{S},$$

and thus the eigenvalues of the cross Gramian are invariant.

The interesting point here is that the "reverse" cross Gramian $\mathscr{O}\mathscr{R}$ stays unchanged under any invertible linear transformation as

$$\hat{\mathscr{O}}\hat{\mathscr{R}} = \mathscr{O}\mathscr{R},$$

where, by definition, is given by

$$\mathscr{O}\mathscr{R} : \mathbb{L}^2\left(\mathbb{R}^p, (-\infty, 0)\right) \to \mathbb{L}^2\left(\mathbb{R}^q, (0, \infty)\right),$$

and

$$\boldsymbol{y}(t) = \mathscr{O}\mathscr{R}\boldsymbol{u} = \boldsymbol{C}\int_{-\infty}^{0} e^{\boldsymbol{A}(t-\tau)}\boldsymbol{B}\boldsymbol{u}\left(\tau\right)\, d\tau, \quad t > 0,$$

which is exactly the past-input-to-future-output map: known as the Hankel operator (see, e.g., [9]).

Under balanced truncation, the Hankel operator changes however. The question is how much the output changes when a balanced truncation is carried out. To answer this question, without loss of generality, let us assume that the system (22.1) is already balanced, and thus both observability and controllability Gramians are equal to a diagonal matrix $\boldsymbol{\Sigma}$. We would like to bound the worst-case relative error

$$\sup_{\boldsymbol{u}} \frac{\|\boldsymbol{y}-\boldsymbol{y}_r\|_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))}}{\|\boldsymbol{u}\|_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}} = \sup_{\boldsymbol{u}} \frac{\|(\mathscr{O}\mathscr{R}-\mathscr{O}_r\mathscr{R}_r)\,\boldsymbol{u}\|_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))}}{\|\boldsymbol{u}\|_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}} = \|\mathscr{O}\mathscr{R}-\mathscr{O}_r\mathscr{R}_r\|$$

Do not know how to estimate this value here. Come back later if necessary.

22.5 A new look at balanced truncation and its application to linear and nonlinear systems

In this section, we first derive a new general framework for observability and controllability. We then apply for the system (22.1) and show that the concept of observability and reachability in section 22.2 and section 22.3 are the special case of the general framework. We then apply the system to two nonlinear systems. We will also discuss how to apply the framework for general input and output energies. We conclude the section with a discussion on a balanced truncation method based on the proposed framework. For the simplicity of the exposition, we assume that all functions under consideration are sufficiently smooth so that their required Fréchet derivatives are well-defined.

22.5.1 General setting

We begin with a general setting. Let $h : \mathbb{X} \ni x \mapsto y = h(x) \in \mathbb{Y}$. We are interested in finding a direction $\hat{x}$ in $\mathbb{X}$ along which y is most sensitive. There are many ways to address this question (see, e.g., [**?**]). Here, we follow [30] to find global sensitive directions on an average sense based on Fréchet derivative (see Chapter 9). To begin, recall the Fréchet derivative with respect to x is denoted as $\mathscr{D}_x h(x)$ which is a linear and continuous mapping from $\mathbb{X}$ to $\mathbb{Y}$. The action of $\mathscr{D}_x h(x)$ in any direction z is denoted as $\mathscr{D}_x h(x)\, z$. Since Fréchet derivative $\mathscr{D}_x h(x)$ depends on the direction z linearly, we can assume that z is a unit vector. *We define the most sensitive direction $\hat{x}$ is the one along which $\mathscr{D}_x h(x)$ amplify the largest*, i.e.,

$$\hat{x} = \arg\max_{\|z\|_{\mathbb{X}}=1} \|\mathscr{D}_x h(x)\, z\|_{\mathbb{Y}}\,.$$

Recall from Remark 5.5 that this is a valid definition as $\mathscr{D}_x h(x)$ is a linear and continuous mapping. Equivalently,

$$\begin{aligned}\hat{x} = \arg\max_{\|z\|_{\mathbb{X}}=1} \|\mathscr{D}_x h(x)\, z\|_{\mathbb{Y}}^2 &= \arg\max_{\|z\|_{\mathbb{X}}=1} (\mathscr{D}_x h(x)\, z, \mathscr{D}_x h(x)\, z)_{\mathbb{Y}} \\ &= \arg\max_{\|z\|_{\mathbb{X}}=1} \left(z, (\mathscr{D}_x h(x))^* \mathscr{D}_x h(x)\, z\right)_{\mathbb{X}},\end{aligned}$$

that is, $\hat{x}$ is an eigenfunction associated with the largest eigenvalue of the self-adjoint operator $\left(\mathscr{D}_x h\left(x\right)\right)^* \mathscr{D}_x h\left(x\right)$. As a result, the above sensitive direction $\hat{x}$ is a local definition as $\mathscr{D}_x h\left(x\right)$ depends on x. In particular, $\hat{x}$ could be the most sensitive direction at a point x but maybe not at another point. This is an issue for methods in which sensitive directions are important. We resolve this by defining globally sensitive directions in an average sense. To that end, let $\mu\left(x\right)$ be a probability measure on $\mathbb{X}$. **The globally sensitive direction $\hat{x}$ is defined as**

$$\hat{x} = \arg \max_{\left\|z\right\|_{\mathbb{X}}=1} \left(z, \mathbb{E}_\mu \left[\left(\mathscr{D}_x h\left(x\right)\right)^* \mathscr{D}_x h\left(x\right)\right] z\right)_{\mathbb{X}}, \tag{22.7}$$

where the expectation is defined as

$$\mathscr{C} := \mathbb{E}_\mu \left[\left(\mathscr{D}_x h\left(x\right)\right)^* \mathscr{D}_x h\left(x\right)\right] := \int_{\mathbb{X}} \left(\mathscr{D}_x h\left(x\right)\right)^* \mathscr{D}_x h\left(x\right) \, d\mu\left(x\right). \tag{22.8}$$

As can be seen, $\mathscr{C}$ is a self-adjoint operator on $\mathbb{X}$ and it is the average of all $\left(\mathscr{D}_x h\left(x\right)\right)^* \mathscr{D}_x h\left(x\right)$ under the probability measure μ. From Chapter 20, we see that $\mathscr{C}$ is nothing more than the average Gauss-Newton Hessian of $\frac{1}{2}\left\|h\left(x\right)\right\|_{\mathbb{Y}}^2$. As a result, $\hat{x}$ defined in (22.7) is a global sensitivity and it is an eigenfunction associated with the largest eigenvalue of average Gauss-Newton Hessian $\mathscr{C}$. The eigenfunctions are sensitivity directions of $h\left(x\right)$ and the eigenvalues allow us to rank eigenfunctions from the most to least sensitive ones, assuming the number of non-zero eigenvalues is countable and the eigenspace associated with an eigenvalue is finite-dimensional. This is guaranteed when h is a compact mapping. The theoretical details on how to exploit this information for dimensional reduction can be found in [30]. In the following, we shall see that observability and reachability gramians derived in (22.2) and (22.5) are special cases of *the sensitivity covariance operator* $\mathscr{C}$. More importantly, the unified setting that we propose in this section provides a unified balanced truncation approach for both linear and nonlinear systems.

22.5.2 A new look at the observability gramian

We now apply the general setting for the observable map $\mathscr{O}$ to derive the observability gramian. In this case, $\mathbb{X} = \mathbb{R}^n$, $\mathbb{Y} = \mathbb{L}^2\left(\mathbb{R}^q, (0,\infty)\right)$ and $h : \mathbb{R}^n \ni \boldsymbol{x}_0 \mapsto h\left(\boldsymbol{x}_0\right) := \boldsymbol{y}\left(\boldsymbol{x}_0\right) = \mathscr{O}\boldsymbol{x}_0 = \boldsymbol{C}e^{\boldsymbol{A}t}\boldsymbol{x}_0 \in \mathbb{L}^2\left(\mathbb{R}^q, (0,\infty)\right)$. Note that $\boldsymbol{y}$, and hence h, is also a function of time t, but we ignore this for the simplicity of the exposition. We next define the new notion of observability based on sensitivity. To that end, we note from section 22.2 that the more observable state is the one produces more output norm

Definition 22.3 (Observability). A unit vector $\hat{\boldsymbol{z}} \in \mathbb{R}^n$ is said to be locally observable at $\boldsymbol{x}_0$ if the magnitude of Fréchet derivative of the output $\boldsymbol{y}$ at $\boldsymbol{x}_0$ along direction $\hat{\boldsymbol{z}}$, $\|D_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\,\hat{\boldsymbol{z}}\|^2_{\mathbb{L}^2(\mathbb{R}^q,(0,\infty))} = \left(\hat{\boldsymbol{z}}, D_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)^* D_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\,\hat{\boldsymbol{z}}\right)_{\mathbb{R}^n}$, is non-zero. That is, an observable state $\hat{\boldsymbol{z}}$ is a direction along which the output $\boldsymbol{y}$ is sensitive. A unit vector $\hat{\boldsymbol{z}} \in \mathbb{R}^n$ is said to be globally observable if on average it is a sensitive direction, i.e., $\left(\hat{\boldsymbol{z}}, \mathbb{E}_{\mu(\boldsymbol{x}_0)}\left[D_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)^* D_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\right]\hat{\boldsymbol{z}}\right)_{\mathbb{R}^n} \neq 0$.

How is this definition related to Definition 22.1? To answer this question, we recall the Taylor expansion up to the first-order term [**?**]

$$\boldsymbol{y}(\boldsymbol{x}_0 + \varepsilon\hat{\boldsymbol{z}}) \approx \boldsymbol{y}(\boldsymbol{x}_0) + \varepsilon\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\,\hat{\boldsymbol{z}},$$

which shows that if $\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\,\hat{\boldsymbol{z}} \neq \theta$, then moving along the direction $\hat{\boldsymbol{z}}$ changes the value of $\boldsymbol{y}$. It also shows that the larger $\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\,\hat{\boldsymbol{z}}$ is, the more $\boldsymbol{y}$ varies. Since moving along $\hat{\boldsymbol{z}}$ induces changes in the output $\boldsymbol{y}$, it is observable in that sense.

As in section 22.2 and section 22.4 we would like to order the observable states for model reduction purposes. In particular, we are interested in the most observable states. From Example 9.2, the Fréchet derivative $\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)$ is given as

$$\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0) = \boldsymbol{C}e^{\boldsymbol{A}t} = \mathscr{O}, \quad \text{and} \quad \left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\right)^* = \mathscr{O}^*.$$

We see that, for the LTI system (22.1), $\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0) = \boldsymbol{C}e^{\boldsymbol{A}t}$ is independent of $\boldsymbol{x}_0$. Now assume the probability measure μ has the probability density $\pi(\boldsymbol{x})$ with respect to the Lebesgue measure on the state space $\mathbb{R}^n$, i.e.,

$$d\mu(\boldsymbol{x}) := \pi(\boldsymbol{x})\,d\boldsymbol{x}.$$

The observable states are then the eigenvectors of

$$\mathscr{C} = \mathbb{E}_{\mu(\boldsymbol{x}_0)}\left[\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\right)^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)\right] = \mathbb{E}_{\pi(\boldsymbol{x}_0)}\left[\mathscr{O}^*\mathscr{O}\right] = \mathscr{O}^*\mathscr{O},$$

where we have used the fact that $\mathscr{O}$ is independent of $\boldsymbol{x}_0$ and $\pi(\boldsymbol{x})\,d\boldsymbol{x}$ is a probability measure on $\mathbb{R}^n$. Thus, we have shown:

- The observabilty gramian $\mathscr{O}^*\mathscr{O}$ in (22.2) is a special case of the sensitivity covariance operator $\mathscr{C}$ in (22.8).
- Globally observable states are directions along which the output $\boldsymbol{y}$ is sensitive on average and they are eigenvectors of the sensitivity covariance matrix $\mathscr{C}$. In particular, the most observable states are eigenvectors associated with the largest eigenvalue of $\mathscr{C}$.
- Eigenvectors with larger eigenvalues of $\mathscr{C}$ are easier to observe.

22.5.3 A new look at the reachability gramian

Recall from section 22.3 that for each initial state $\boldsymbol{x}_0$, we are interested in finding the input with minimum energy that steers the system (22.1) from zero state at negative infinity time to $\boldsymbol{x}_0$ at time $t = 0$. The solution was given in (22.4) and was derived based on the linearity of the reachable operator $\mathscr{R}$ and the Lagrangian approach. We then showed that the initial condition $\boldsymbol{x}_0$ that gives rise to the input with the smallest energy is any eigenvector associated with the largest eigenvalue of the reachability gramian. In the following we will show that the reachability gramian in (22.5) is a special case of the sensitivity covariance operator $\mathscr{C}$ in (22.8) corresponding to a particular setting of $\mathbb{X}, \mathbb{Y}$ and h. To make our approach valid for nonlinear systems, we revisit the problem of finding a minimum norm control input $\boldsymbol{u}(t)$ associated with an initial condition $\boldsymbol{x}_0$ using a more general approach. The problem at hand is

$$\min_{\boldsymbol{u}\in\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} \frac{1}{2}\left\|\boldsymbol{u}\right\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))},$$

subject to the following constraints

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \quad \text{and} \quad \boldsymbol{x}(-\infty) = \boldsymbol{0}.$$

The constraint $\boldsymbol{x}(-\infty) = \boldsymbol{0}$ implies that the equilibrium $\boldsymbol{0}$ is asymptotically stable not only for the system in the absence of the control input $\boldsymbol{u}$ but also for the system with the optimal control.

Following the Lagrangian multiplier Theorem 9.3, we consider the Lagrangian

$$L(\boldsymbol{u},\boldsymbol{x},\boldsymbol{z},\boldsymbol{a},\boldsymbol{b}) := \frac{1}{2}\left\|\boldsymbol{u}\right\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + \left(\boldsymbol{z}, \frac{d\boldsymbol{x}}{dt} - \boldsymbol{A}\boldsymbol{x} - \boldsymbol{B}\boldsymbol{u}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + \boldsymbol{a}^T\left[\boldsymbol{x}(0) - \boldsymbol{x}_0\right] + \boldsymbol{b}^T\boldsymbol{x}(-\infty).$$

We can now apply Theorem 9.3 to derive the (first order) optimality condition. The vanishing of the Fréchet derivatives of L with respect to $\boldsymbol{z}$, $\boldsymbol{a}$, and $\boldsymbol{b}$ gives back the constraints. Setting to 0 the Fréchet derivative of L with respect to $\boldsymbol{x}$ and integrating by parts once we have

$$-\int_{-\infty}^{0} \boldsymbol{p}^T\left(\frac{d\boldsymbol{z}}{dt} + \boldsymbol{A}^T\boldsymbol{z}\right) dt + \boldsymbol{p}(0)^T\left[\boldsymbol{a} + \boldsymbol{z}(0)\right] + \boldsymbol{p}(-\infty)^T\left(\boldsymbol{b} - \boldsymbol{z}(-\infty)\right) = 0$$

for $\boldsymbol{p} \in \mathbb{L}^2(\mathbb{R}^n,(-\infty,0))$. Restricting to the dense subspace $\mathcal{C}_0^\infty(\mathbb{R}^n,(-\infty,0))$ of $\mathbb{L}^2(\mathbb{R}^n,(-\infty,0))$, we have

$$-\int_{-\infty}^{0} \boldsymbol{p}^T \left(\frac{d\boldsymbol{z}}{dt} + \boldsymbol{A}^T \boldsymbol{z} \right) dt = 0,$$

and thus

$$\frac{d\boldsymbol{z}}{dt} + \boldsymbol{A}^T \boldsymbol{z} = \boldsymbol{0}.$$

Next taking $\boldsymbol{p}(t)$ is such that $\boldsymbol{p}(0) = \boldsymbol{0}$ and then $\boldsymbol{p}(-\infty) = \boldsymbol{0}$ gives $\boldsymbol{z}(-\infty) = \boldsymbol{b}$ and $\boldsymbol{z}(0) = -\boldsymbol{a}$. It follows that $\boldsymbol{z}(0)$ and $\boldsymbol{z}(-\infty)$ are finite and the Lagrangian multipliers $\boldsymbol{a}$ and $\boldsymbol{b}$ can be found once $\boldsymbol{z}(t)$ is solved for.

Finally, setting to vanish the Fréchet derivative of L with respect to $\boldsymbol{u}$ we obtain

$$(\boldsymbol{u}, \boldsymbol{v})_{\mathbb{L}^2(\mathbb{R}^n,(-\infty,0))} - \left(\boldsymbol{B}^T \boldsymbol{z}, \boldsymbol{v} \right)_{\mathbb{L}^2(\mathbb{R}^n,(-\infty,0))} = 0,$$

for any $\boldsymbol{v} \in \mathbb{L}^2(\mathbb{R}^n, (-\infty, 0))$. It follows that

$$\boldsymbol{u}(t) = \boldsymbol{B}^T \boldsymbol{z}(t).$$

In summary, the first-order optimality conditions are given as

$$\frac{d\boldsymbol{x}}{dt} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{u}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \quad \text{and} \quad \boldsymbol{x}(-\infty) = \boldsymbol{0}, \tag{22.9a}$$

$$\frac{d\boldsymbol{z}}{dt} = -\boldsymbol{A}^T \boldsymbol{z}, \tag{22.9b}$$

$$\boldsymbol{u} = \boldsymbol{B}^T \boldsymbol{z}. \tag{22.9c}$$

Solving (22.9a) for $\boldsymbol{x}(0)$ gives

$$\boldsymbol{x}_0 = \boldsymbol{x}(0) = \int_{-\infty}^{0} e^{-\boldsymbol{A}t} \boldsymbol{B}\boldsymbol{u}(t)\, dt. \tag{22.10}$$

Next, solving (22.9b) for $\boldsymbol{z}(t)$ gives

$$\boldsymbol{z}(t) = e^{-\boldsymbol{A}^T t} \boldsymbol{z}(0).$$

Substituting $\boldsymbol{z}(t)$ into (22.9c) we obtain

$$\boldsymbol{u}(t) = \boldsymbol{B}^T e^{-\boldsymbol{A}^T t} \boldsymbol{z}(0),$$

which is then substituted in (22.10) to give

$$\boldsymbol{x}_0 = \int_{-\infty}^{0} e^{-\boldsymbol{A}t} \boldsymbol{B}\boldsymbol{B}^T e^{-\boldsymbol{A}^T t}\, dt\, \boldsymbol{z}(0) = \mathscr{R}\mathscr{R}^* \boldsymbol{z}(0).$$

Similar to section 22.3, with the complete reachability assumption, i.e., $\dim(\mathsf{R}(\mathscr{R})) = n$, $\mathscr{R}\mathscr{R}^*$ is invertible. Combining the last equations and the definition of $\mathscr{R}^*$ in section 22.3 yields the optimal control input $\boldsymbol{u}$ (now a

function of the initial state $\boldsymbol{x}_0$)

$$\boldsymbol{u}(\boldsymbol{x}_0, t) = \mathscr{R}^* \left(\mathscr{R}\mathscr{R}^*\right)^{-1} \boldsymbol{x}_0, \tag{22.11}$$

which, as expected, is the same results as in (22.4).*Note that the optimal control input* $\boldsymbol{u}(t)$ *is a function of both time* t *and the initial state* $\boldsymbol{x}_0$*, but for notational convenience we typically write* $\boldsymbol{u}(\boldsymbol{x}_0)$ *with the dependency on time* t *implicitly understood.*

We now apply the general framework in subsection 22.5.1 with $\mathbb{X} = \mathbb{R}^n$, $\mathbb{Y} = \mathbb{L}^2\left(\mathbb{R}^p, (-\infty, 0)\right)$, and $h\left(\boldsymbol{x}_0\right) := \boldsymbol{u}\left(\boldsymbol{x}_0\right) = \mathscr{R}^* \left(\mathscr{R}\mathscr{R}^*\right)^{-1} \boldsymbol{x}_0$. Note that $\boldsymbol{u}$, and hence h, is a also a function of time (since $\mathscr{R}^*$ is), and this is understood implicitly for the simplicity of notations. Similar to subsection 22.5.2, we assume that the probability measure μ has the probability density $\pi\left(\boldsymbol{x}\right)$ with respect to the Lebesgue measure on the state space $\mathbb{R}^n$. We are in the position to define reachability based on Fréchet derivative of $\boldsymbol{u}$.

Definition 22.4 (Reachability). A unit vector $\hat{\boldsymbol{z}} \in \mathbb{R}^n$ is said to be locally reachable at $\boldsymbol{x}_0$ if the magnitude of Fréchet derivative of the optimal control input $\boldsymbol{u}$ at $\boldsymbol{x}_0$ along direction $\hat{\boldsymbol{z}}$, $\left\|D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\hat{\boldsymbol{z}}\right\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} = \left(\hat{\boldsymbol{z}}, D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)^* D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\hat{\boldsymbol{z}}\right)_{\mathbb{R}^n}$, is non-zero. That is, a reachable state $\hat{\boldsymbol{z}}$ is a direction along which the optimal control input $\boldsymbol{u}$ is sensitive. A unit vector $\hat{\boldsymbol{z}} \in \mathbb{R}^n$ is said to be globally reachable if on average it is a sensitive direction, i.e., $\left(\hat{\boldsymbol{z}}, \mathbb{E}_{\mu(\boldsymbol{x}_0)}\left[D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)^* D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\right]\hat{\boldsymbol{z}}\right)_{\mathbb{R}^n} \neq 0$.

The relationship between Definition 22.4 and Definition 22.2 can be seen through Taylor expansion up to the first-order term

$$\boldsymbol{u}\left(\boldsymbol{x}_0 + \varepsilon\hat{\boldsymbol{z}}\right) \approx \boldsymbol{u}\left(\boldsymbol{x}_0\right) + \varepsilon\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\hat{\boldsymbol{z}}.$$

Indeed, as long as $D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\hat{\boldsymbol{z}} \neq \theta$, moving along $\hat{\boldsymbol{z}}$ is associated with a non-zero control $\boldsymbol{u}\left(\boldsymbol{x}_0 + \varepsilon\hat{\boldsymbol{z}}\right)$, and hence the existence of the control $\boldsymbol{u}$ at $\boldsymbol{x}_0 + \varepsilon\hat{\boldsymbol{z}}$ provided the control $\boldsymbol{u}$ at $\boldsymbol{x}_0$ exists. As in section 22.3 we are interested in the globally most reachable direction $\hat{\boldsymbol{x}}$. It is $\hat{\boldsymbol{z}}$ for which $\left(\hat{\boldsymbol{z}}, \mathbb{E}_{\mu(\boldsymbol{x}_0)}\left[D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)^* D_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\right]\hat{\boldsymbol{z}}\right)_{\mathbb{R}^n}$ is smallest since moving along such a direction induces smallest changes in $\boldsymbol{u}$. Back to our intuition on $\boldsymbol{u}$ as cost and $\boldsymbol{x}_0$ as gain, gaining along such a $\hat{\boldsymbol{z}}$ on average requires the smallest changes in the cost.

From Example 9.2, the Fréchet derivative $\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)$ is given as

$$\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right) = \mathscr{R}^*\left(\mathscr{R}\mathscr{R}^*\right)^{-1}, \quad \text{and} \quad \left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\right)^* = \left(\mathscr{R}\mathscr{R}^*\right)^{-1}\mathscr{R},$$

which are independent of $\boldsymbol{x}_0$. *The globally reachable states are then the eigenfunctions corresponding to non-zero eigenvalue of the self-adjoint operator* $\mathscr{C}$ *in* (22.8). Furthermore, eigenfunctions associated with smaller eigenvalues are more reachable. For the LTI system (22.1), $\mathscr{C}$ is an $n \times n$ invertible symmetric matrix given as

$$\begin{aligned}\mathscr{C} &= \mathbb{E}_{\mu(\boldsymbol{x}_0)}\left[\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\right)^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\right] \\ &= \mathbb{E}_{\pi(\boldsymbol{x}_0)}\left[\left(\mathscr{R}\mathscr{R}^*\right)^{-1}\mathscr{R}\mathscr{R}^*\left(\mathscr{R}\mathscr{R}^*\right)^{-1}\right] = \left(\mathscr{R}\mathscr{R}^*\right)^{-1}.\end{aligned}$$

As can be seen, the more reachable states are equivalently the eigenvectors of $\mathscr{R}\mathscr{R}^*$ associated with larger eigenvalues. In particular, those associated with the largest eigenvalue of $\mathscr{R}\mathscr{R}^*$ are the most reachable. Thus, we have shown that our new notion of reachability reduces to the standard one for the LTI system (22.1), and the reachability gramian is rediscovered as the inverse of the sensitivity covariance matrix in this case.

In summary:

- The reachability gramian $\mathscr{R}\mathscr{R}^*$ in (22.5) is a special case of the inverse of the sensitivity covariance operator $\mathscr{C}$ in (22.8).
- Globally reachable states are directions along which the input $\boldsymbol{u}$ is sensitive on average and they are eigenvectors of the reachability gramian. In particular, the most reachable states are eigenvectors associated with the largest eigenvalue of the reachability gramian $\mathscr{C}^{-1}$ (assuming the inverse exists).
- Eigenvectors with larger eigenvalues of $\mathscr{C}^{-1}$ are easier to reach.

22.5.4 Computing new observability gramian for nonlinear systems

We now consider a nonlinear system of the following form

$$\begin{cases}\dfrac{d\boldsymbol{x}}{dt} &= f\left(\boldsymbol{x}\right) + g\left(\boldsymbol{x}\right)\boldsymbol{u}, \\ \boldsymbol{y} &= h\left(\boldsymbol{x}\right),\end{cases} \tag{22.12}$$

where $f : \mathbb{R}^n \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}^{n\times p}$, and $h : \mathbb{R}^n \to \mathbb{R}^q$ are vector-valued, matrix-valued functions, and vector-valued functions, respectively. We assume f, g and h so that their partial derivatives up to second order are meaningful. Furthermore, we assume f, g and h are such that $\boldsymbol{y} \in \mathbb{L}^2\left(\mathbb{R}^q, (0,\infty)\right)$ for any $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $\boldsymbol{u} \in \mathbb{L}^2\left(\mathbb{R}^p, (-\infty, 0)\right)$. We also assume that $\mathbf{0}$ is an equilibrium of the nonlinear system in the absence of $\boldsymbol{u}$ and it is asymptotically stable (see Definition 21.2). We are interested in constructing a balanced truncation method for the nonlinear system (22.12) and we are going to apply the unified setting in subsection 22.5.1 to accomplish our goal.

As shown in section 22.4, balanced truncation requires **three ingredients**: i) observable states are defined and ranked, ii) reachable states are defined and ranked, and iii) a transformation through which a new state is equally observable and reachable. The first and second ingredients were essentially already discussed in subsection 22.5.2 and subsection 22.5.3. Indeed, for the

observability gramian, the only place where we used linearity of (22.1) was in the last step of showing $\mathscr{C} = \mathscr{O}^*\mathscr{O}$. Ignoring this step, we conclude that observable states are eigenvectors of the sensitivity covariance matrix (with the new notation $\mathscr{C}_O$)

$$\mathscr{C}_O = \mathbb{E}_{\pi(\boldsymbol{x}_0)}\left[\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^*\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right], \tag{22.13}$$

where $\boldsymbol{y}\left(\boldsymbol{x}_0\right) := \boldsymbol{y}\left(\boldsymbol{x}\left(\boldsymbol{x}_0\right)\right) \in \mathbb{L}^2\left(\mathbb{R}^q, (0,\infty)\right)$ is the output of (22.12) starting from $\boldsymbol{x}_0$ at $t = 0$ with zero input $\boldsymbol{u}$. Eigenvectors of $\mathscr{C}_O$ associated with larger eigenvalues are more observable.

Definition 22.5 (Observability gramian). Let π be a probability density function on $\mathbb{R}^n$. The observability gramian $\mathcal{G}_O$ for the system (22.12) is defined as $\mathscr{C}_O$ in (22.13). The system (22.12) is said completely observable if $\mathcal{G}_O$ is invertable.

A few observations are in order. First, forming $\mathscr{C}_O \in \mathbb{R}^{n\times n}$ could be computationally extensive as we need to evaluate the expectation of $\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^*\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)$ under the probability density $\pi\left(\boldsymbol{x}_0\right)$. Clearly, this is intractable for nonlinear h and/or nontrivial π. One popular approach is to approximate the expectation with some Monte Carlo approach [69, 103, 122, 80, 38, 125, 124, 38, 101, 48, 110, 64, 20]. This amounts to evaluating $\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^*\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)$ for each sample $\boldsymbol{x}_0$. Second, note that $\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right) : \mathbb{R}^n \to \mathbb{L}^2\left(\mathbb{R}^q, (0,\infty)\right)$ and $\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^* : \mathbb{L}^2\left(\mathbb{R}^q, (0,\infty)\right) \to \mathbb{R}^n$, and thus evaluating $\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^*\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)$ for each sample $\boldsymbol{x}_0$ is in turn an integration from 0 to ∞ due to the inner product in $\mathbb{L}^2\left(\mathbb{R}^q, (0,\infty)\right)$. In practice, we may have to truncate the integral at some large time T or deploy some transformation and then use some special quadrature [68, 120, 67, 43].

Note that for the purpose of balanced truncation, we only need to compute the dominant part of the spectrum of $\mathscr{C}_O$. To that end, we can deploy any methods [**?**] that require only the matrix-vector products. We begin with the following result.

Proposition 22.1. *Given $\boldsymbol{x}_0 \in \mathbb{R}^n$, $\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^*\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)$ is the Gauss-Newton Hessian of the function $J_O : \mathbb{R}^n \to \mathbb{R}$ defined as*

$$J_O\left(\boldsymbol{x}_0\right) := \frac{1}{2}\left\|\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} = \frac{1}{2}\left\|h\left(\boldsymbol{x}_0\right)\right\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}. \tag{22.14}$$

where $\boldsymbol{x}$ is the solution of the nonlinear dynamical system (22.12) with $\boldsymbol{u} = \boldsymbol{0}$ and initial condition $\boldsymbol{x}\left(0\right) = \boldsymbol{x}_0$.

Proof. Let $\boldsymbol{H}_{GN}\left(\boldsymbol{x}_0\right) \in \mathbb{R}^{n\times n}$ be the Gauss-Newton Hessian of $J_O\left(\boldsymbol{x}_0\right)$. For any $\boldsymbol{w}, \boldsymbol{z} \in \mathbb{R}^n$, we have

$$\begin{aligned}\left(\boldsymbol{w}, \boldsymbol{H}_{GN}\left(\boldsymbol{x}_0\right)\boldsymbol{z}\right)_{\mathbb{R}^n} &= \left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\boldsymbol{w}, \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\boldsymbol{z}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}\\ &= \left(\boldsymbol{w}, \left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^*\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\boldsymbol{z}\right)_{\mathbb{R}^n},\end{aligned}$$

which shows that $\boldsymbol{H}_{GN}(\boldsymbol{x}_0) = (\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0))^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}(\boldsymbol{x}_0)$.

We next present a procedure for computing the product of the Gauss-Newton Hessian J_O with an arbitrary vector $\hat{\boldsymbol{x}}_0$ in two steps: i) computing the gradient $\nabla J_O(\boldsymbol{x}_0)$; and then ii) computing the product the Gauss-Newton hessian with an arbitrary vector $\hat{\boldsymbol{x}}_0$: $(\nabla J_O(\boldsymbol{x}_0))^* \nabla J_O(\boldsymbol{x}_0)\hat{\boldsymbol{x}}_0$.

22.5.4.1 Computing $\nabla J_O(\boldsymbol{x}_0)$

Since we will provide step-by-step detail in section 22.3 for the reachability gramian, which is a more complicated case, we only summarize the result here and leave the detailed derivation as an exercise for the readers (see Problem 22.3).

Proposition 22.2. *The gradient $\nabla J_O(\boldsymbol{x}_0)$ is given as*

$$\nabla J_O(\boldsymbol{x}_0) = -\boldsymbol{w}(0),$$

where $\boldsymbol{w}(t)$, together with $\boldsymbol{x}(t)$, is the solution of the following system

$$\frac{d\boldsymbol{w}}{dt} = -\nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T \boldsymbol{w} + \nabla_{\boldsymbol{x}} h(\boldsymbol{x})^T h(\boldsymbol{x}), \quad \boldsymbol{w}(\infty) = \boldsymbol{0}, \tag{22.15a}$$

and

$$\frac{d\boldsymbol{x}}{dt} = f(\boldsymbol{x}), \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \tag{22.15b}$$

provided that (22.15) has a unique solution[2] for any given $\boldsymbol{x}_0$.

The solution process for (22.15) is clear: i) solve (22.15b) for $\boldsymbol{x}$ as a function $\boldsymbol{x}_0$, and then ii) solve (22.15a) for $\boldsymbol{w}$ as a function of $\boldsymbol{x}$ and $\boldsymbol{w}(0)$. From $\lim_{t\to\infty} \boldsymbol{w}(t) = \boldsymbol{0}$, we can solve for $\boldsymbol{w}(0)$ as a function of $\boldsymbol{x}_0$. Note that $-\boldsymbol{w}(0)$ is exactly the gradient of J_O at $\boldsymbol{x}_0$ if we need to evaluate the gradient.

22.5.4.2 Computing $(\nabla J_O(\boldsymbol{x}_0))^* \nabla J_O(\boldsymbol{x}_0)\hat{\boldsymbol{x}}_0$

Let $\hat{\boldsymbol{x}}_0$ be an arbitrary vector in $\mathbb{R}^n$. The task at hand is to compute the matrix-vector product $(\nabla J_O(\boldsymbol{x}_0))^* \nabla J_O(\boldsymbol{x}_0)\hat{\boldsymbol{x}}_0$. Since $(\nabla J_O(\boldsymbol{x}_0))^* \nabla J_O(\boldsymbol{x}_0)$ is the Gauss-Newton Hessian of $J_O(\boldsymbol{x}_0)$, we can apply the results in Chapter 20 to derive the procedure to compute $\nabla J_O(\boldsymbol{x}_0)^* \nabla J_O(\boldsymbol{x}_0)\hat{\boldsymbol{x}}_0$ exactly. For the simplicity of the exposition, let us denote the directional (Fréchet) derivative of any quantity $(\cdot)$ at $\boldsymbol{x}_0$ in the direction $\hat{\boldsymbol{x}}_0$ as $\hat{(\cdot)}$. For example, $\hat{\boldsymbol{x}}$

[2] A sufficient condition for the uniqueness of the solution is that $f(\boldsymbol{x})$ and $h(\boldsymbol{x})$, together with their first partial derivatives, are globally Lipschitz continuous. Weaker conditions can be imposed [95], but this is not the interest of the book.

is the directional derivative of $\boldsymbol{x}$ at $\boldsymbol{x}_0$ in direction $\hat{\boldsymbol{x}}_0$. Clearly, the directional derivative of $\boldsymbol{x}_0$ along the direction $\hat{\boldsymbol{x}}_0$ is exactly $\hat{\boldsymbol{x}}_0$. To simply the notations, let us define F^j as the jth row of $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$, and H^j as the jth row of $\nabla_{\boldsymbol{x}} h(\boldsymbol{x})$.

It follows from Proposition 22.2 that

$$\nabla J_O(\boldsymbol{x}_0)^* \nabla J_O(\boldsymbol{x}_0) \hat{\boldsymbol{x}}_0 = -\hat{\boldsymbol{w}}(0),$$

where $\hat{\boldsymbol{w}}(t)$, coupled with $\hat{\boldsymbol{x}}$, is the solution of the following system

$$\frac{d\hat{\boldsymbol{w}}}{dt} = -\nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T \hat{\boldsymbol{w}} - \nabla_{\boldsymbol{x}} h(\boldsymbol{x})^T \nabla_{\boldsymbol{x}} h(\boldsymbol{x}) \hat{\boldsymbol{x}}, \quad \hat{\boldsymbol{w}}(\infty) = \mathbf{0}, \tag{22.16a}$$

and

$$\frac{d\hat{\boldsymbol{x}}}{dt} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \hat{\boldsymbol{x}}, \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0, \tag{22.16b}$$

where $\boldsymbol{x}$ and $\boldsymbol{w}$ have already been solved from (22.15) as we discussed at the end of the previous section.

Computing $\nabla J_O(\boldsymbol{x}_0)^* \nabla J_O(\boldsymbol{x}_0) \hat{\boldsymbol{x}}_0$ thus amounts to computing $\hat{\boldsymbol{w}}(0)$. This can be carried out in two steps as we did in solving (22.15). First, we solve for $\hat{\boldsymbol{x}}$ as a function of $\hat{\boldsymbol{x}}_0$ from (22.16b). Second, we solve for $\hat{\boldsymbol{w}}$ as a function of $\hat{\boldsymbol{w}}(0)$ and $\hat{\boldsymbol{x}}$. By passing to the limit $\lim_{t\to\infty} \hat{\boldsymbol{w}}(t) = \mathbf{0}$, we can compute $\hat{\boldsymbol{w}}(0)$ as a function of $\hat{\boldsymbol{x}}_0$.

22.5.5 Computing new reachability gramian for nonlinear systems

For the reachability gramian, if we do not exploit the linearity of the system (22.1) that allows for the explicit formula for the Fréchet derivative of the input $\boldsymbol{u}$ with respect to $\boldsymbol{x}_0$, the following conclusion still holds: reachable states are eigenvectors corresponding to non-zero eigenvalues of the sensitivity covariance matrix (with new notation $\mathscr{C}_R$)

$$\mathscr{C}_R = \mathbb{E}_{\pi(\boldsymbol{x}_0)} \left[\left(\mathscr{D}_{\boldsymbol{x}_0} \boldsymbol{u}(\boldsymbol{x}_0) \right)^* \mathscr{D}_{\boldsymbol{x}_0} \boldsymbol{u}(\boldsymbol{x}_0) \right], \tag{22.17}$$

and the states that are easier to reach are those eigenvectors of $\mathscr{C}_R$ associated with smaller eigenvalues.

Definition 22.6 (Reachability gramian). Let π be a probability density function on $\mathbb{R}^n$. The system (22.12) is completely reachable if $\mathscr{C}_R$ in (22.17) is invertible. Suppose that the system (22.12) is completely reachable, then the reachability gramian $\mathcal{G}_R$ for the system (22.12) is defined as $\mathscr{C}_R^{-1}$.

Remark 22.1. When $\mathscr{C}_R$ in (22.17) is not invertible, we define the reachability gramian $\mathcal{G}_R$ as the pseudo-inverse of $\mathscr{C}_R$.

The issues with evaluating the expectation under the probability density π in $\mathbb{R}^n$ and the integration for semi-infinite interval $(0,\infty)$ that we have seen in computing $\mathscr{C}_O$ are also pertinent to the evaluation of $\mathscr{C}_R$. Note that we also do not know if $\mathscr{C}_R$ is invertible in general. Furthermore, evaluating $\mathscr{C}_R$ gives rise to another issue. While evaluating the Fréchet derivative for $\boldsymbol{y}$ is straightforward, it is much harder for $\boldsymbol{u}$. The reason is that for nonlinear system (22.12), we cannot express $\boldsymbol{u}$ as a function of $\boldsymbol{x}_0$ explicitly because the nonlinear counterpart of the first-order conditions (22.9) is not solvable analytically in general, as we shall see. The Fréchet derivative is thus also implicit. In the next two sections, we address the question of how to compute $\boldsymbol{u}(\boldsymbol{x}_0)$, the Fréchet derivative $\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0)$, and the adjoint $(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0))^*$.

22.5.5.1 Finding $\boldsymbol{u}(\boldsymbol{x}_0)$

To compute $\boldsymbol{u} = \boldsymbol{u}(\boldsymbol{x}_0)$, in fact $\boldsymbol{u}(\boldsymbol{x}_0, t)$ but we ignore t for simplicity, we need to revisit the problem of finding a minimum norm input $\boldsymbol{u}$ that drives the system from $\mathbf{0}$ at infinity to $\boldsymbol{x}_0$ at $t = 0$, but now for the nonlinear system (22.12):

$$\min_{\boldsymbol{u}\in\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} \frac{1}{2}\|\boldsymbol{u}\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))}, \tag{22.18a}$$

subject to the following constraints

$$\frac{d\boldsymbol{x}}{dt} = f(\boldsymbol{x}) + g(\boldsymbol{x})\boldsymbol{u}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0 \quad \text{and} \quad \boldsymbol{x}(-\infty) = \mathbf{0}, \tag{22.18b}$$

with $\boldsymbol{x}_0$ being in an asymptotically stable neighborhood[3] $\mathcal{N}$ of $\mathbf{0}$. It follows that an optimal input control $\boldsymbol{u}$, if exists, ensures that $\mathbf{0}$ is also asymptotically stable for (22.18b).

To solve (22.18) we follow the same Lagrangian approach that we did for the linear system (22.1) above. In particular, define the Lagrangian as

$$L(\boldsymbol{u},\boldsymbol{x},\boldsymbol{z},\boldsymbol{a},\boldsymbol{b}) := \frac{1}{2}\|\boldsymbol{u}\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + \left(\boldsymbol{z}, \frac{d\boldsymbol{x}}{dt} - f(\boldsymbol{x}) - g(\boldsymbol{x})\boldsymbol{u}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + \boldsymbol{a}^T(\boldsymbol{x}(0) - \boldsymbol{x}_0) + \boldsymbol{b}^T\boldsymbol{x}(-\infty).$$

Again, the vanishing of the Fréchet derivatives of L with respect to $\boldsymbol{z}$, $\boldsymbol{a}$, and $\boldsymbol{b}$ gives back the constraints. Setting the Fréchet derivative of L with respect to $\boldsymbol{x}$ along an arbitrary direction $\boldsymbol{p}$ to zero we have

[3] Here, an asymptotically stable neighborhood $\mathcal{N}$ means: if the system (22.18b) starts from $\boldsymbol{x}_0 \in \mathcal{N}$ at time $t = 0$, then $\boldsymbol{x}(t) \to \mathbf{0}$ as $t \to \infty$ in the absence of $\boldsymbol{u}$.

$$\left(\boldsymbol{z}, \frac{d\boldsymbol{p}}{dt} - \nabla_{\boldsymbol{x}} f(\boldsymbol{x})\boldsymbol{p} - \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})\boldsymbol{p}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + \boldsymbol{a}^T\boldsymbol{p}(0) + \boldsymbol{b}^T\boldsymbol{p}(-\infty) = 0, \quad \forall \boldsymbol{p} \in \mathbb{L}^2(\mathbb{R}^p,(-\infty,0)),$$

where g^i be the ith column of g and $\boldsymbol{u}_i$ is the ith component of $\boldsymbol{u}$. After integrating by parts we obtain

$$-\left(\boldsymbol{p}, \frac{d\boldsymbol{z}}{dt} + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T\boldsymbol{z} + \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T\boldsymbol{z}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + \boldsymbol{p}(0)^T[\boldsymbol{z}(0) + \boldsymbol{a}] + \boldsymbol{p}(-\infty)^T(\boldsymbol{b} - \boldsymbol{z}(-\infty)) = 0, \quad \forall \boldsymbol{p} \in \mathbb{L}^2(\mathbb{R}^p,(-\infty,0)),$$

which, after choosing different cases for $\boldsymbol{p}$ similar to the linear system in subsection 22.5.3, implies

$$\frac{d\boldsymbol{z}}{dt} + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T\boldsymbol{z} + \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T\boldsymbol{z} = \boldsymbol{0}. \tag{22.19}$$

Note that $\boldsymbol{a} = -\boldsymbol{z}(0)$ and $\boldsymbol{b} = \boldsymbol{z}(-\infty)$ are part of the conclusion, and this implies that $\boldsymbol{z}(0)$ and $\boldsymbol{z}(-\infty)$ are finite and that the Lagrangian multiplier $\boldsymbol{a}$ and $\boldsymbol{b}$ can be computed once $\boldsymbol{z}(t)$ is solved for from (22.19). Next setting the Fréchet derivative of L with respect to $\boldsymbol{u}$ to zero we have

$$(\boldsymbol{u}, \boldsymbol{v})_{\mathbb{L}^2(\mathbb{R}^n,(-\infty,0))} - \left(g(\boldsymbol{x})^T\boldsymbol{z}, \boldsymbol{v}\right)_{\mathbb{L}^2(\mathbb{R}^n,(-\infty,0))} = 0,$$

for any $\boldsymbol{v} \in \mathbb{L}^2(\mathbb{R}^n,(-\infty,0))$. It follows that

$$\boldsymbol{u} = g(\boldsymbol{x})^T\boldsymbol{z}.$$

In summary, we have found the first-order optimality conditions (also called the first-order forward system in the next section) for the optimal control problem (22.18).

Lemma 22.1. *An optimal solution $\boldsymbol{u}(\boldsymbol{x}_0,t) \in \mathbb{L}^2(\mathbb{R}^p,(-\infty,0))$ of the optimal control problem (22.18), if exists, is a solution of the following system:*

$$\frac{d\boldsymbol{x}}{dt} = f(\boldsymbol{x}) + g(\boldsymbol{x})\boldsymbol{u}, \quad \boldsymbol{x}(0) = \boldsymbol{x}_0, \quad \textit{and} \quad \boldsymbol{x}(-\infty) = \boldsymbol{0} \tag{22.20a}$$

$$\frac{d\boldsymbol{z}}{dt} = -\nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T\boldsymbol{z} - \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T\boldsymbol{z}, \tag{22.20b}$$

$$\boldsymbol{u} = g(\boldsymbol{x})^T\boldsymbol{z}. \tag{22.20c}$$

The system of algebraic and ordinary differential equations (22.20) allows us to solve for the control input $\boldsymbol{u}$ as a function of $\boldsymbol{x}_0$. Assume that there

is a solution for (22.20) and one way to verify this is through the implicit function Theorem 9.1 similar to the discussion after Corollary 9.2, but this is not the interest of this chapter.[4] We now describe an analytical procedure to compute $\boldsymbol{u}(\boldsymbol{x}_0, t)$ provided that all the steps can be computed analytically (e.g. in closed-form expressions).

F1) Substituting (22.20c) into (22.20a), we can the solve (22.20a) for $\boldsymbol{x}$ as a function of $\boldsymbol{x}_0$ and $\boldsymbol{z}$, i.e., $\boldsymbol{x}(t) = F(\boldsymbol{x}_0, \boldsymbol{z}(\tau), t)$. Note that τ is a dummy time variable indicating $\boldsymbol{z}$ being a function of time, but $\boldsymbol{x}(t)$ is not a function of τ as it is integrated out when computing $\boldsymbol{x}(t)$.

F2) Next, substituting (22.20c) and $\boldsymbol{x}(t) = F(\boldsymbol{x}_0, \boldsymbol{z}(\tau), t)$ into (22.20b), we can solve for $\boldsymbol{z}$ as a function of $\boldsymbol{z}(0)$ and $\boldsymbol{x}_0$, i.e., $\boldsymbol{z}(\tau) = G(\boldsymbol{z}(0), \boldsymbol{x}_0, \tau)$.

F3) It follows that $\boldsymbol{x}(t) = F(\boldsymbol{x}_0, G(\boldsymbol{z}(0), \boldsymbol{x}_0, \tau), t)$. Now, setting $t = 0$ we have $\boldsymbol{x}_0 = \boldsymbol{x}(0) = F(\boldsymbol{x}_0, G(\boldsymbol{z}(0), \boldsymbol{x}_0, \tau), 0)$ and this allows us to solve for $\boldsymbol{z}(0)$ as a function of $\boldsymbol{x}_0$, i.e., $\boldsymbol{z}(0) = H(\boldsymbol{x}_0)$.

F4) Finally substituting $\boldsymbol{z}(\tau) = G(H(\boldsymbol{x}_0), \boldsymbol{x}_0, \tau)$ into (22.20c) we have

$$\boldsymbol{u}(\boldsymbol{x}_0, t) = g(F(\boldsymbol{x}_0, G(H(\boldsymbol{x}_0), \boldsymbol{x}_0, \tau), t))^T G(H(\boldsymbol{x}_0), \boldsymbol{x}_0, t), \quad (22.21)$$

which is the optimal control input depending only on $\boldsymbol{x}_0$.

Note that the above analytical procedure is limited to simple scenarios for the nonlinear systems (22.12) in which all the steps can be solved in closed-form expressions. When this is infeasible, we need to resort to some numerical optimization method to solve the optimal control problem (22.18) for $\boldsymbol{u}(\boldsymbol{x}_0, t)$. Since this is not of interest to this chapter, we refer the readers to some standard literature [119, 118, 29, 61].

Next, we need to compute the sensitivity covariance matrix $\mathscr{C}_R$ in (22.17). Exact computation for $\mathscr{C}_R$ is generally not possible even when $\pi(\boldsymbol{x}_0)$ is a Gaussian probability density due to the nonlinearity of $(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0))^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0)$ in terms of $\boldsymbol{x}_0$. In practice, we deploy some Monte Carlo approach to approximate the expectation in (22.17). Thus, computing $\mathscr{C}_R$ amounts to computing $(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0))^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0)$ at different values of $\boldsymbol{x}_0$ (sampled from $\pi(\boldsymbol{x}_0)$), and then taking average. It is therefore sufficient to be able to evaluate $(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0))^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0)$ for a given $\boldsymbol{x}_0$. It turns out that it is the Gauss-Newton Hessian of a "simple" function as we now show.

Corollary 22.1. *Given $\boldsymbol{x}_0 \in \mathbb{R}^n$, $(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0))^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}(\boldsymbol{x}_0)$ is the Gauss-Newton Hessian of the function $J_R : \mathbb{R}^n \to \mathbb{R}$ defined as*

$$J_R(\boldsymbol{x}_0) := \frac{1}{2} \|\boldsymbol{u}(\boldsymbol{x}_0)\|^2_{\mathbb{L}^2(\mathbb{R}^p, (-\infty, 0))}, \quad (22.22)$$

[4] Another approach is to substitute (22.20c) into (22.20a) and (22.20b). We can then impose a sufficient condition on $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ such that the system of coupled ordinary differential equations (ODE) (22.20a) and (22.20b) have a unique solution $[\boldsymbol{x}, \boldsymbol{z}]$, starting from an initial condition, using ODE theories (see, e.g., [95]). For example, the requirement that $f(\boldsymbol{x})$ and $g(\boldsymbol{x})$ are differentiable and their derivatives are globally Lipschitz continuous is sufficient.

where $\boldsymbol{u}$ *is an optimal solution of* (22.18).

Proof. Since $\boldsymbol{u}$ is an optimal solution of (22.18), it is a solution of the first order forward system (22.20). By $\boldsymbol{u}(\boldsymbol{x}_0)$ we mean a solution of (22.18) as a function of $\boldsymbol{x}_0$, for example, from the procedure Item **F1)**—Item **F4)**. The rest of the proof is similar to the proof of Proposition 22.1, and thus is omitted.

Corollary 22.1 shows that J_R in (22.22) depends on $\boldsymbol{x}_0$ implicitly through the solution $\boldsymbol{u}$ of (22.20) that induces a smallest J for a given $\boldsymbol{x}_0$. From now on we assume that $\boldsymbol{u}(\boldsymbol{x}_0)$ is such a solution.

In the following, we present a procedure for computing the Gauss-Newton Hessian of J_R for a given $\boldsymbol{x}_0$ in two steps: i) compute the gradient $\nabla J_R(\boldsymbol{x}_0)$ and then ii) compute the product of the Gauss-Newton Hessian $\nabla J_R(\boldsymbol{x}_0)^* \nabla J_R(\boldsymbol{x}_0)$ with an arbitrary vector $\hat{\boldsymbol{x}}_0 \in \mathbb{R}^n$. To be discussed below, for dimensional reduction purposes, we do not need to compute the full Gauss-Newton Hessian, but only the relevant part of its eigenspaces. In order to achieve such a purpose, we need only the product of the Gauss-Newton Hessian with arbitrary vectors, because this is what efficient matrix-free algorithms [**?**] need.

22.5.5.2 Finding the gradient $\nabla J_R(\boldsymbol{x}_0)$

This section presents an approach to compute the reduced gradient approach to compute the gradient $\nabla J_R(\boldsymbol{x}_0)$. Let us define F^j as the jth row of $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$, G^{ij} as the jth row of $\nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})$, and the following Lagrangian

$$
\begin{aligned}
L := \frac{1}{2}\|\boldsymbol{u}\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} + \left(\boldsymbol{y}, \frac{d\boldsymbol{x}}{dt} - f(\boldsymbol{x}) - g(\boldsymbol{x})\boldsymbol{u}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} \\
+(\boldsymbol{\alpha}, \boldsymbol{x}(0) - \boldsymbol{x}_0)_{\mathbb{R}^n} + \left(\boldsymbol{w}, \frac{d\boldsymbol{z}}{dt} + \nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T \boldsymbol{z} + \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T \boldsymbol{z}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} \\
+ (\boldsymbol{\beta}, \boldsymbol{z}(-\infty))_{\mathbb{R}^n} + \left(\boldsymbol{v}, \boldsymbol{u} - g(\boldsymbol{x})^T \boldsymbol{z}\right)_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))},
\end{aligned}
$$

where $\boldsymbol{y}, \boldsymbol{w}, \boldsymbol{v} \in \mathbb{L}^2(\mathbb{R}^p, (-\infty, 0))$, and $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^n$ are the Lagrangian multipliers.

Proposition 22.3. *The gradient* $\nabla J(\boldsymbol{x}_0)$ *is given by*

$$\nabla J_R(\boldsymbol{x}_0) = \boldsymbol{y}(0),$$

where $\boldsymbol{y}(t)$*, together with* $\boldsymbol{w}(t)$ *and* $\boldsymbol{v}(t)$*, satisfies the following "first order adjoint" system:*

$$\frac{d\boldsymbol{y}}{dt} = -\nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T \boldsymbol{y} - \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T \boldsymbol{y} - \sum_{i=1}^{p} \boldsymbol{v}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T \boldsymbol{z} + \sum_{j=1}^{n} \boldsymbol{z}_j \left(\nabla_{\boldsymbol{x}} F^j\right)^T \boldsymbol{w} + \sum_{i=1,j=1}^{p,n} \boldsymbol{u}_i \boldsymbol{z}_j \left(\nabla_{\boldsymbol{x}} G^{ij}\right)^T \boldsymbol{w}, \quad (22.23a)$$

$$\frac{d\boldsymbol{w}}{dt} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \boldsymbol{w} + \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x}) \boldsymbol{w} + g(\boldsymbol{x}) \boldsymbol{v}, \quad \boldsymbol{w}(0) = \boldsymbol{0}, \quad \text{and } \boldsymbol{w}(-\infty) = 0, \quad (22.23b)$$

and

$$\boldsymbol{u} + \boldsymbol{v} - g(\boldsymbol{x})^T \boldsymbol{y} - \boldsymbol{\lambda} = \boldsymbol{0}, \quad (22.23c)$$

where $\boldsymbol{\lambda} \in \mathbb{L}^2(\mathbb{R}^n, (-\infty, 0))$ *with the* i*th component defined as*

$$\boldsymbol{\lambda}_i = \left(\boldsymbol{z}, \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x}) \boldsymbol{w}\right)_{\mathbb{R}^n},$$

and $\boldsymbol{x}, \boldsymbol{z}$ *and* $\boldsymbol{u}$ *satisfy (22.20).*

Proof. The proof is the same as Lemma 9.4 and the derivation of (22.20) in subsection 22.5.5.1, and thus is left as an exercise (see Problem 22.4).

What remains is how to solve the first order adjoint system (22.23) for $\boldsymbol{y}, \boldsymbol{w}$ and $\boldsymbol{v}$ provided that $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{u}$ have been already solved for from the first order forward system (22.20). Similar to (22.20), we now describe an analytical approach to obtain a closed form expression for the solution for (22.23).

A1) Substituting $\boldsymbol{v}$ from (22.23c) into (22.23b), we can the solve (22.23b) for $\boldsymbol{w}$ as a function of $\boldsymbol{y}$, i.e., $\boldsymbol{w}(t) = \mathcal{F}(\boldsymbol{x}_0, \boldsymbol{y}(\tau), t)$. Note that τ is a dummy time variable indicating $\boldsymbol{y}$ being a function of time, but $\boldsymbol{w}(t)$ is not a function of τ as it is integrated out when computing $\boldsymbol{w}(t)$.

A2) Next, substituting $\boldsymbol{v}$ from (22.23c) and $\boldsymbol{w}(t) = \mathcal{F}(\boldsymbol{x}_0, \boldsymbol{y}(\tau), t)$ into (22.23a), we can solve for $\boldsymbol{y}$ as a function of $\boldsymbol{y}(0)$ and $\boldsymbol{x}_0$, i.e., $\boldsymbol{y}(\tau) = \mathcal{G}(\boldsymbol{y}(0), \boldsymbol{x}_0, \tau)$.

A3) It follows that $\boldsymbol{w}(t) = \mathcal{F}(\boldsymbol{x}_0, \mathcal{G}(\boldsymbol{y}(0), \boldsymbol{x}_0, \tau), t)$. Now, setting $t = 0$ we have $\boldsymbol{w}(0) = \boldsymbol{0} = \mathcal{F}(\boldsymbol{x}_0, \mathcal{G}(\boldsymbol{y}(0), \boldsymbol{x}_0, \tau), 0)$ and this allows us to solve for $\boldsymbol{y}(0)$ as a function of $\boldsymbol{x}_0$, i.e., $\boldsymbol{y}(0) = \mathcal{H}(\boldsymbol{x}_0)$.

A4) Finally substituting $\boldsymbol{y}(\tau) = \mathcal{G}(\mathcal{H}(\boldsymbol{x}_0), \boldsymbol{x}_0, \tau)$ into (22.23c) we have

$$\boldsymbol{v}(\boldsymbol{x}_0, t) = -\boldsymbol{u} + g(\boldsymbol{x})^T \mathcal{G}(\mathcal{H}(\boldsymbol{x}_0), \boldsymbol{x}_0, t) + \boldsymbol{\lambda}, \quad (22.24)$$

where the ith component of $\boldsymbol{\lambda}$, as a function of $\boldsymbol{x}_0$ and t, is given as

$$\boldsymbol{\lambda}_i(\boldsymbol{x}_0, t) = \left(\boldsymbol{z}, \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x}) \mathcal{F}(\boldsymbol{x}_0, \mathcal{G}(\mathcal{H}(\boldsymbol{x}_0), \boldsymbol{x}_0, \tau), t)\right)_{\mathbb{R}^n}.$$

Provided that all steps in the procedure Item **A1)**—Item **A4)** is analytically tractable, we have $\boldsymbol{w}$, $\boldsymbol{y}$, and $\boldsymbol{v}$ as a function of $\boldsymbol{x}_0$ (and t) at this point. In addition, the gradient ∇J depends on $\boldsymbol{x}_0$ implicitly through

$$\nabla J_R(\boldsymbol{x}_0) = \boldsymbol{y}(0) = \mathcal{H}(\boldsymbol{x}_0).$$

22.5.5.3 Finding the Gauss-Newton Hessian-vector product $\nabla J_R(\boldsymbol{x}_0)^* \nabla J_R(\boldsymbol{x}_0) \hat{\boldsymbol{x}}_0$

Let $\hat{\boldsymbol{x}}_0$ be an arbitrary vector in $\mathbb{R}^n$. Recall that the product of the Gauss-Newton Hessian of J_R with $\hat{\boldsymbol{x}}_0$ is part of the product of the Hessian of J_R with $\hat{\boldsymbol{x}}_0$. As a result, we can apply the results in Chapter 20 to derive the procedure to compute $\nabla J_R(\boldsymbol{x}_0)^* \nabla J_R(\boldsymbol{x}_0) \hat{\boldsymbol{x}}_0$ exactly. To that end, let us denote the directional (Fréchet) derivative of any quantity $(\cdot)$ at $\boldsymbol{x}_0$ in an arbitrary direction $\hat{\boldsymbol{x}}_0$ as $\hat{(\cdot)}$. For example, $\hat{\boldsymbol{x}}$ is the directional derivative of $\boldsymbol{x}$ at $\boldsymbol{x}_0$ in direction $\hat{\boldsymbol{x}}_0$. Note that the directional derivative of $\boldsymbol{x}_0$ along the direction $\hat{\boldsymbol{x}}_0$ is exactly $\hat{\boldsymbol{x}}_0$. It is important to point out that at this point Proposition 22.3 establishes that the reduced gradient $\nabla J_R(\boldsymbol{x}_0)$ has been computed where $\boldsymbol{x}, \boldsymbol{z}$ and $\boldsymbol{u}$ depend on $\boldsymbol{x}_0$ via the first order forward equations (22.20) and $\boldsymbol{y}, \boldsymbol{w}$ and $\boldsymbol{v}$ depend on $\boldsymbol{x}_0$ via the first order adjoint equations (22.23). It follows from Proposition 22.3 that

$$\nabla J_R(\boldsymbol{x}_0)^* \nabla J_R(\boldsymbol{x}_0) \hat{\boldsymbol{x}}_0 = \hat{\boldsymbol{y}}(0),$$

where $\hat{\boldsymbol{y}}(t)$, together with $\hat{\boldsymbol{w}}(t)$ and $\hat{\boldsymbol{v}}(t)$, is obtained from the (simplified) second order adjoint system

$$\frac{d\hat{\boldsymbol{y}}}{dt} = -\nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T \hat{\boldsymbol{y}} - \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T \hat{\boldsymbol{y}} - \sum_{i=1}^{p} \hat{\boldsymbol{v}}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T \boldsymbol{z} + \sum_{j=1}^{n} \boldsymbol{z}_j \left(\nabla_{\boldsymbol{x}} F^j\right)^T \hat{\boldsymbol{w}} + \sum_{i=1,j=1}^{p,n} \boldsymbol{u}_i \boldsymbol{z}_j \left(\nabla_{\boldsymbol{x}} G^{ij}\right)^T \hat{\boldsymbol{w}}, \quad (22.25a)$$

$$\frac{d\hat{\boldsymbol{w}}}{dt} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) \hat{\boldsymbol{w}} + \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x}) \hat{\boldsymbol{w}} + g(\boldsymbol{x}) \hat{\boldsymbol{v}}, \qquad \hat{\boldsymbol{w}}(0) = \boldsymbol{0}, \quad \hat{\boldsymbol{w}}(-\infty) = \boldsymbol{0}, \quad (22.25b)$$

and

$$\hat{\boldsymbol{u}} + \hat{\boldsymbol{v}} - g(\boldsymbol{x})^T \hat{\boldsymbol{y}} - \hat{\boldsymbol{\lambda}} = \boldsymbol{0}, \quad (22.25c)$$

where $\hat{\boldsymbol{\lambda}} \in \mathbb{L}^2(\mathbb{R}^n, (-\infty, 0))$ with the ith component defined as

$$\hat{\boldsymbol{\lambda}}_i = \left(\boldsymbol{z}, \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x}) \hat{\boldsymbol{w}}\right)_{\mathbb{R}^n}.$$

Note that in order to solve (22.25) we need to first solve for $\hat{\boldsymbol{u}}(t)$, together with $\hat{\boldsymbol{x}}$ and $\hat{\boldsymbol{z}}$, from the second order forward system

$$\frac{d\hat{\boldsymbol{x}}}{dt} = \nabla f(\boldsymbol{x})\hat{\boldsymbol{x}} + \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})\,\hat{\boldsymbol{x}} + g(\boldsymbol{x})\hat{\boldsymbol{u}}, \quad \hat{\boldsymbol{x}}(0) = \hat{\boldsymbol{x}}_0, \quad \hat{\boldsymbol{x}}(-\infty) = 0, \tag{22.26a}$$

$$\frac{d\hat{\boldsymbol{z}}}{dt} = -\nabla_{\boldsymbol{x}} f(\boldsymbol{x})^T \hat{\boldsymbol{z}} - \sum_{j=1}^{n} \boldsymbol{z}_j \nabla_{\boldsymbol{x}} F^j \hat{\boldsymbol{x}} - \sum_{i=1}^{p} \boldsymbol{u}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T \hat{\boldsymbol{z}} - \sum_{i=1,j=1}^{p,n} \boldsymbol{u}_i \boldsymbol{z}_j \left(\nabla_{\boldsymbol{x}} G^{ij}\right)^T \hat{\boldsymbol{x}} - \sum_{i=1}^{p} \hat{\boldsymbol{u}}_i \nabla_{\boldsymbol{x}} g^i(\boldsymbol{x})^T \boldsymbol{z}, \tag{22.26b}$$

$$\hat{\boldsymbol{u}} = \sum_{j=1}^{n} \boldsymbol{z}_j \nabla_{\boldsymbol{x}} g_j(\boldsymbol{x})\,\hat{\boldsymbol{x}} + g(\boldsymbol{x})^T \hat{\boldsymbol{z}}. \tag{22.26c}$$

where g_j is the jth row of $g(\boldsymbol{x})$.

Suppose that it is analytically (in closed form expression) tractable, we can solve the second order forward system (22.26) for $\hat{\boldsymbol{u}}(t)$, $\hat{\boldsymbol{x}}$, and $\hat{\boldsymbol{z}}$ just like we did for the first order forward equation (22.20) following the procedure Item **F1)**—Item **F4)**. Similarly, solving the second order adjoint system (22.25) can be carried out by following a similar procedure as we did in Item **A1)**—Item **A4)**.

Remark 22.2. We would like to point out that the result of section 22.5 still holds for any finite time horizon $\mathcal{T}$ by simply replacing ∞ with $\mathcal{T}$.

22.5.6 Balanced truncation

Recall from Definition 22.5 and Definition 22.6 that the observability and reachability gramians matrices in $\mathbb{R}^{n\times n}$ defined as

$$\mathcal{G}_O = \mathscr{C}_O, \quad \text{and} \quad \mathcal{G}_R = \mathscr{C}_R^{-1}.$$

We also recall that states align with eigenvectors of observability gramian $\mathcal{G}_O$ (reachability gramian $\mathcal{G}_R$ respectively) associated with larger eigenvalues are more observable (more reachable respectively). Similar to section 22.4, we would like to transform the system (22.12) such that a transformed state is equally observable and reachable. Such a transformation must simultaneously diagonalizes observability and reachability Gramians so that the transformed observability and reachability Gramians are not only diagonal but also equal to each other. It turns out that with our definitions of observability and reachability gramians, the procedure for balancing transformation is the same.

Lemma 22.2. *Let π be a probability density function on $\mathbb{R}^n$ and suppose the system (22.12) is completely reachable and observable. Suppose the system (22.12) and the optimal control problem (22.18) have unique solutions. Consider the following two-step invertible transformation:*

1. *Let the eigenvalue decomposition of $\mathcal{G}_O$ be $\mathcal{G}_O = \boldsymbol{U}\boldsymbol{\Lambda}\boldsymbol{U}^T$. Transform the system (22.12) via $\boldsymbol{x} = \boldsymbol{S}\hat{\boldsymbol{x}}$, where $\boldsymbol{S} = \boldsymbol{U}\Lambda^{-1/2}$.*
2. *Let $\hat{\mathcal{G}}_R$ be the new reachability gramian associated with the new state $\hat{\boldsymbol{x}}$ and let the eigenvalue decomposition of $\hat{\mathcal{G}}_R$ be $\hat{\mathcal{G}}_R = \boldsymbol{V}\boldsymbol{\Sigma}^2\boldsymbol{V}^T$. Transform the system once more time with $\hat{\boldsymbol{x}} = \boldsymbol{T}\tilde{\boldsymbol{x}}$ where $\boldsymbol{T} = \boldsymbol{V}\boldsymbol{\Sigma}^{1/2}$.*

Then the observability and reachability gramians $\tilde{\mathcal{G}}_O$ and $\tilde{\mathcal{G}}_R$ associated with $\tilde{\boldsymbol{x}}$ are equal to the diagonal matrix $\boldsymbol{\Sigma}$, and thus any component of $\tilde{\boldsymbol{x}}$ is equally observable and reachable.

Proof. It is sufficient to see how the first transformation $\boldsymbol{x} = \boldsymbol{S}\hat{\boldsymbol{x}}$ changes the gramians where $\boldsymbol{S} \in \mathbb{R}^{n\times n}$ is an invertiable matrix. Let $\hat{\mathcal{G}}_O$ be the gramian associated with $\hat{\boldsymbol{x}}$. The system (22.12) is transformed to

$$\begin{cases} \dfrac{d\hat{\boldsymbol{x}}}{dt} & = \boldsymbol{S}^{-1} f\left(\boldsymbol{S}\hat{\boldsymbol{x}}\right) + \boldsymbol{S}^{-1} g\left(\boldsymbol{S}\hat{\boldsymbol{x}}\right)\boldsymbol{u}, \quad \hat{\boldsymbol{x}}\left(0\right) = \boldsymbol{S}^{-1}\boldsymbol{x}_0 \\ \boldsymbol{y} & = h\left(\boldsymbol{S}\hat{\boldsymbol{x}}\right), \end{cases} \tag{22.27}$$

For the observability gramian, with $\boldsymbol{u} = 0$ and the uniqueness of the solution of (22.12), and hence uniqueness of the solution $\hat{\boldsymbol{x}}$ of (22.27), we can write

$$\boldsymbol{y} = \boldsymbol{y}\left(\boldsymbol{x}_0\right) = \boldsymbol{y}\left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right),$$

which implies

$$\mathscr{D}_{\hat{\boldsymbol{x}}_0}\boldsymbol{y}\left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right) = \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\boldsymbol{S}.$$

It follows that

$$\left(\mathscr{D}_{\hat{\boldsymbol{x}}_0}\boldsymbol{y}\left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right)\right)^* \mathscr{D}_{\hat{\boldsymbol{x}}_0}\boldsymbol{y}\left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right) = \boldsymbol{S}^T\left(\mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\right)^* \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{y}\left(\boldsymbol{x}_0\right)\boldsymbol{S}.$$

We conclude that $\hat{\mathcal{G}}_O = \boldsymbol{S}^T\mathcal{G}_O\boldsymbol{S}$. For the reachability gramian, a similar argument shows that

$$\mathscr{D}_{\hat{\boldsymbol{x}}_0}\boldsymbol{u}\left(\boldsymbol{S}\hat{\boldsymbol{x}}_0\right) = \mathscr{D}_{\boldsymbol{x}_0}\boldsymbol{u}\left(\boldsymbol{x}_0\right)\boldsymbol{S},$$

and thus

$$\hat{\mathcal{G}}_R = \boldsymbol{S}^{-1}\mathcal{G}_R\boldsymbol{S}^{-T}.$$

Thus if we apply the proposed transformations, the conclusion of the proposition follows.

Definition 22.7 (Balanced system). The system (22.12) is called balanced its observability and reachability gramians are diagonal and equal.

Now suppose the system (22.12) is balanced with both observability and reachability gramians are equal to a diagonal matrix $\boldsymbol{\Sigma}$. Let $\sigma_i := \boldsymbol{\Sigma}\left(i,i\right)$.

We can order the diagonal terms of $\boldsymbol{\Sigma}$ such that $\sigma_i \geq \sigma_j \geq 0$ for all $i \leq j$ and $i, j \in \{1, \ldots, n\}$. Let $\boldsymbol{x}_r$ be the first r components of $\boldsymbol{x}$ and they are corresponding to the first r largest eigenvalues of the gramians $\mathcal{G}_R = \mathcal{G}_O = \boldsymbol{\Sigma}$. The remaining part of $\boldsymbol{x}$, corresponding to smaller eigenvalues, is denoted as $\boldsymbol{x}_t$.

Clearly, at this point, any component of the state $\boldsymbol{x}$ is equally observable and reachable. Thus, it makes sense to remove the portion $\boldsymbol{x}_t$. This can be done by truncating the portions of the system, particularly $f(x), g(\boldsymbol{x})$ and $h(\boldsymbol{x})$, corresponding to $\boldsymbol{x}_t$. Unlike the linear counterpart discussed in section 22.4, the remaining parts of these quantities could be still a function of $\boldsymbol{x}_t$. See below for the discussion of how to remove $\boldsymbol{x}_t$ entirely by marginalizing out $\boldsymbol{x}_t$ (based on the active subspace view point). I STILL DO NOT KNOW HOW TO CHARACTERIZE THE ERROR AS IN THE ACTIVE SUBSPACE APPROACHE. PERHAPS, THAT IS IMPOSSIBLE. COME BACK TO THIS LATER.

The question that we would like to address now is how to derive the reduced dynamical system for $\boldsymbol{x}_r$ and the "reduced" output $\boldsymbol{y}_r$. We define the reduced output via conditional distribution

$$\boldsymbol{y}_r(\boldsymbol{x}_r) := \mathbb{E}_{\pi(\boldsymbol{x}_t|\boldsymbol{x}_r)}[\boldsymbol{y}(\boldsymbol{x})],$$

where $\pi(\boldsymbol{x}_t|\boldsymbol{x}_r)$ is the conditional distribution of $\boldsymbol{x}_t$ given $\boldsymbol{x}_r$ and is given as

$$\pi(\boldsymbol{x}_t|\boldsymbol{x}_r) = \frac{\pi(\boldsymbol{x})}{\pi(\boldsymbol{x}_r)},$$

where $\pi(\boldsymbol{x}_r)$ is the marginal distribution of $\boldsymbol{x}_r$.

Problems

Problem 22.1. Show that if the system (22.1) is completely observable, then the observability Gramian $\mathscr{O}^*\mathscr{O} : \mathbb{R}^n \to \mathbb{R}^n$ is invertible. Similarly, show that if the system (22.1) is completely reachable, then the reachability Gramian $\mathscr{R}\mathscr{R}^* : \mathbb{R}^n \to \mathbb{R}^n$ is invertible.

Hint: Suppose $\boldsymbol{x}$ is such that $\mathscr{R}\mathscr{R}^*\boldsymbol{x} = \theta$. Then $\|\mathscr{R}^*\boldsymbol{x}\|^2_{\mathbb{L}^2(\mathbb{R}^p,(-\infty,0))} = 0$, which implies that $\boldsymbol{x} \in \mathsf{N}(\mathscr{R}^*) = \mathsf{R}(\mathscr{R})^\perp = (\mathbb{R}^n)^\perp = \{\theta\}$. Thus, $\boldsymbol{x} = \theta$ and hence the matrix $\mathscr{R}\mathscr{R}^*$ is invertible.

Problem 22.2. Consider the linear transformation $\boldsymbol{x} = \boldsymbol{S}\hat{\boldsymbol{x}}$, where $\boldsymbol{S}$ is an invertible matrix. The system (22.1) is transformed into the new system

$$\begin{cases} \dfrac{d\hat{\boldsymbol{x}}}{dt} &= \hat{\boldsymbol{A}}\hat{\boldsymbol{x}} + \hat{\boldsymbol{B}}\boldsymbol{u}, \\ \boldsymbol{y} &= \hat{\boldsymbol{C}}\hat{\boldsymbol{x}}, \end{cases}$$

where $\hat{\boldsymbol{A}} = \boldsymbol{S}^{-1}\boldsymbol{A}\boldsymbol{S}$, $\hat{\boldsymbol{B}} = \boldsymbol{S}^{-1}\boldsymbol{B}$, and $\hat{\boldsymbol{C}} = \boldsymbol{C}\boldsymbol{S}$. Show that the observability Gramian $\hat{\mathscr{O}}^*\hat{\mathscr{O}}$ and reachability Gramian $\hat{\mathscr{R}}\hat{\mathscr{R}}^*$ of the new system are given by

$$\hat{\mathscr{O}}^*\hat{\mathscr{O}} = \boldsymbol{S}^T\mathscr{O}^*\mathscr{O}\boldsymbol{S}, \text{ and } \hat{\mathscr{R}}\hat{\mathscr{R}}^* = \boldsymbol{S}^{-1}\mathscr{R}\mathscr{R}^*\boldsymbol{S}^{-T},$$

and thus

$$\hat{\mathscr{R}}\hat{\mathscr{O}} = \boldsymbol{S}^{-1}\mathscr{R}\mathscr{O}\boldsymbol{S}.$$

Problem 22.3. Prove Proposition 22.2.

Problem 22.4. Prove Proposition 22.3.

Chapter 23
An adjoint approach for error correction and *a posteriori* error estimation

Abstract

- error correction from Mike Giles
- A posterior error estimation from Ranacher. The prerequisite for this chapter is

  1. Frechét derivative in Chapter 9
  2.

23.1 Taylor expansion in function spaces

This section derives Taylor expansion in function spaces with integral remainder. This will be useful for error correction in section 23.4 and a posteriori error estimation in section 23.5. We shall show that Taylor expansions in function spaces are direct consequences of Taylor expansions in one dimension. For concreteness, we consider only Taylor expansion with a specific integral remainder, though the idea can be extended to other remainders in a straightforward manner.

Let $f : \mathbb{R} \ni u \mapsto f(u) \in \mathbb{R}$ be an $(n+1)$ times differentiable function with derivatives $f^{(i)}(u)$, $i = 1, \ldots, n+1$. By the fundamental theorem of calculus, we have

$$f(a+h) = f(a) + \int_0^1 f^{(1)}(a+th)\, h\, dt.$$

Applying the fundamental theorem of calculus again for $f^{(1)}$ gives

$$f(a+h) = f(a) + f^{(1)}(a)\, h + \int_0^1 \int_0^t f^{(2)}(a+sh)\, h^2\, ds\, dt,$$

which, after switching the integrals in the last term, becomes

$$f(a+h) = f(a) + f^{(1)}(a)\,h + \int_0^1 \int_s^1 f^{(2)}(a+sh)\,h^2\,dt\,ds$$
$$= f(a) + f^{(1)}(a)\,h + \int_0^1 f^{(2)}(a+sh)\,(1-s)\,h^2\,ds.$$

By induction, we obtain an nth order Taylor expansion in the following form

$$f(a+h) = \sum_{i=0}^{n} f^{(i)}(a)\,\frac{h^i}{i!} + \frac{1}{n!}\int_0^1 f^{(n+1)}(a+th)\,[(1-t)\,h]^n\,h\,dt \tag{23.1}$$

Let $\mathbb{U}$ be a Hilbert space (the results in this section are also valid for Banach spaces), and consider the functional

$$\mathcal{F} : \mathbb{U} \ni u \mapsto \mathcal{F}(u) \in \mathbb{R}, \tag{23.2}$$

which is assumed to possess Fréchet derivative up to order $(n+1)$. For notational convenience, we use

$$\mathscr{D}\mathcal{F}^n(u; w_1, \ldots, w_n)$$

to denote the nth Fréchet derivative of $\mathcal{F}$ at u along the directions $w_1, \ldots, w_n$, respectively. If $w_1 = \ldots = w_n = w$, we simply write

$$\mathscr{D}\mathcal{F}^n(u; w^n) := \mathscr{D}\mathcal{F}^n(u; w, \ldots, w).$$

We are interested in expanding $\mathcal{F}$ around $u \in \mathbb{U}$ using the one-dimensional Taylor expansion in (23.1). To that end, let us parametrize $\mathcal{F}$ as

$$f(s) := \mathcal{F}(u + sv), \text{ where } \quad s \in \mathbb{R}, \quad v \in \mathbb{U}.$$

Our goal is to apply (23.1) for $f(s)$ with $a = 0$ and $h = 1$. Before doing so, we need to derive the expression for the derivatives $f^{(i)}$, and these can be obtained by the chain rule for Fréchet derivatives (see the footnote[3] in section 9.4). In particular, we can show (see Problem 23.1)

$$f^{(i)}(0) = \mathscr{D}\mathcal{F}^n(u; v^n). \tag{23.3}$$

Putting the results together gives the desired Taylor expansion for F as

$$\mathcal{F}(u+v) = \sum_{i=0}^{n} \frac{\mathscr{D}\mathcal{F}^n(u; v^n)}{i!} + \frac{1}{n!}\int_0^1 \mathscr{D}\mathcal{F}^{n+1}\left(u+tv; v^{n+1}\right)(1-t)^n\,dt \tag{23.4}$$

Clearly, (23.4) reduces to (23.1) when $\mathbb{U} = \mathbb{R}$.

By invoking the mean value theorem for integral, we can rewrite the Taylor expansion in (23.4) with the Cauchy remainder term

$$\mathcal{F}(u+v) = \sum_{i=0}^{n} \frac{\mathscr{D}\mathcal{F}^n(u;v^n)}{i!} + \frac{1}{n!} D\mathcal{F}^{n+1}\left(u+t^*v;v^{n+1}\right)(1-t^*)^n, \quad (23.5)$$

where t^* is some number in $[0,1]$.

23.2 Trapezoidal rule in function spaces

This section derives the trapezoidal rule that we will use in many places in this chapter. As always, our starting point is for functions on the real line $\mathbb{R}$ and for this case, our exposition follows [39] closely. Consider $f : (a,b) \subset \mathbb{R} \ni u \mapsto f(u) \in \mathbb{R}$ be a two-times differentiable function with derivatives $f^{(i)}(u)$, $i = 1, 2$. Suppose that we are interested in approximating the area "under" f over the interval (a,b)

$$I := \int_a^b f(x)\, dx,$$

with the trapezoidal rule

$$I_{\text{trape}} := \frac{b-a}{2}\left[f(b)+f(a)\right],$$

which incurs the error

$$E := I_{\text{trape}} - I = \underbrace{\frac{b-a}{2}\left[f(b)+f(a)\right]} - \int_a^b f(x)\, dx.$$

We are interested in deriving a useful presentation for the error E. The interesting observation[1] about the error E is that it looks like an integration by part result of some integral. Indeed, first, we can write the trapezoidal rule as

$$I_{\text{trape}} = (x-c)\, f(x)\big|_a^b,$$

where c is clearly $\dfrac{b+a}{2}$. Second, the factor $(x-c)$ works out perfectly for the original integral as we can write the error E as

$$E := (x-c)\, f(x)\big|_a^b - \int_a^b (x-c)'\, f(x)\, dx = \int_a^b (x-c) f'(x)\, dx, \quad (23.6)$$

where the prime ' denotes the derivative in x. At this point, if we know that the first derivative $f'(x)$ is uniformly bounded on (a,b) then can immediately

[1] We guess this is what the authors in [39] saw though they didn't provide any intuition on how they came up with the smart idea of "backward" integration by parts.

upper-bound the error E:

$$|E| \leq \|f\|_\infty \frac{(b-a)^2}{2}.$$

On the other hand, if f is twice differentiable, then we can attempt to integrate by parts one more time to obtain

$$E = -\frac{1}{2}\int_a^b (x-c)^2 f''(x)\,dx + \frac{1}{2}(x-c)^2 f'(x)\Big|_a^b.$$

The problem is that we do not know $f'(b)$ nor $f'(a)$. How do we get rid of the boundary terms? The observation is again simple as we can freely add a constant d at an appropriate place without affecting the result:

$$E = -\frac{1}{2}\int_a^b \left[(x-c)^2 + d\right] f''(x)\,dx + \frac{1}{2}\left[(x-c)^2 + d\right] f'(x)\Big|_a^b.$$

Clearly, by choosing $d = -\dfrac{(b-a)^2}{4}$, the boundary term vanishes regardless of what $f'(b)$ and $f'(a)$ are. It follows that

$$\begin{aligned} E = -\frac{1}{2}\int_a^b \left[\left(x - \frac{b+a}{2}\right)^2 - \frac{(b-a)^2}{4}\right] f''(x)\,dx \\ = -\frac{1}{2}\int_a^b (x-a)(x-b) f''(x)\,dx. \end{aligned} \tag{23.7}$$

If we assume that the second derivative f'' is uniformly bounded, then the error E is bounded above as

$$|E| \leq \|f''\|_\infty \frac{(b-a)^3}{12}.$$

Following the same approach in section 23.1, we now lift the one dimensional trapezoidal rule to infinite dimensions. Let $\mathbb{U}$ be a Hilbert space (again, the results in this section are also valid for Banach spaces), and consider the functional

$$\mathcal{G} : \mathbb{U} \ni u \mapsto \mathcal{G}(u) \in \mathbb{R},$$

which is assumed to be twice Fréchet differentiable. Suppose we would like to compute the following integral

$$\int_0^1 \mathcal{G}(u + tv; v)\,dt,$$

with trapezoidal rule. Given u and v, the function $f(t) := \mathcal{G}(u + tv; v)$ is a function on $\mathbb{R}$. Applying the trapezoidal error representation in (23.7) for $f(t)$ gives

$$\begin{aligned}\int_0^1 \mathcal{G}(u + tv; v)\, dt &= \frac{1}{2}\left[\mathcal{G}(u) + \mathcal{G}(u + v)\right] + \frac{1}{2}\int_0^1 t(t-1) f''(t)\, dt \\ &= \frac{1}{2}\left[\mathcal{G}(u) + \mathcal{G}(u + v)\right] + \frac{1}{2}\int_0^1 t(t-1) \mathscr{D}^2\mathcal{G}(u + tv; v, v)\, dt. \quad (23.8)\end{aligned}$$

If $\mathcal{G}$ is only Fréchet differentiable, then we can use the error representation in (23.6) to arrive at

$$\begin{aligned}\int_0^1 \mathcal{G}(u + tv; v)\, dt &= \frac{1}{2}\left[\mathcal{G}(u) + \mathcal{G}(u + v)\right] \\ &\quad - \int_0^1 \left(t - \frac{1}{2}\right) \mathscr{D}\mathcal{G}(u + tv; v)\, dt. \quad (23.9)\end{aligned}$$

23.3 Problem statement

Consider an (possibly nonlinear) operator $\mathscr{A} : \mathbb{U} \ni u \mapsto \mathscr{A}(u) \in \mathbb{V}$, and the following problem: given $f \in \mathbb{V}$, find a solution $u \in \mathbb{U}$ such that

$$\mathscr{A}(u) = f. \quad (23.10)$$

In this chapter, we assume that equation (23.10) has a unique solution u. When $\mathscr{A}$ is a linear operator, the conditions for having a unique solution can be referred to Chapter 15.

For many practical applications, we are not interested in the solution u itself but some quantity of interest as a function of u. Without loss of generality, let us consider the functional

$$J : \mathbb{U} \ni u \mapsto J(u) \in \mathbb{R}.$$

Suppose that solving for u, and hence evaluating $J(u)$, is computationally expensive or impossible. This is typically the case when (23.10) is highly nonlinear (partial differential equations) and/or $\mathbb{U}$ and $\mathbb{V}$ are infinite dimensional spaces. In that case, we resort to some numerical approximation of (23.10):

$$\mathscr{A}_h(u_h) = f_h, \quad (23.11)$$

where h denotes a discretization fidelity (such as mesh size or time steps size in numerical methods). Here, $\mathscr{A}_h$, u_h, and f_h a discretization of $\mathscr{A}, u$, and

f, respectively, and we assume that $u_h \in \mathbb{U}$, $f_h \in \mathbb{V}$, and $\mathscr{A}_h(\cdot) : \mathbb{U} \ni u_h \mapsto \mathscr{A}_h(u_h) \in \mathbb{V}$.

Remark 23.1. Note that the discrete forward equation (23.11) (and the discrete adjoint equation (23.16) in the following) is typically valid only in some subspaces of the original spaces $\mathbb{U}$ and $\mathbb{V}$. In particular, $u_h \in \mathbb{U}_h \subseteq \mathbb{U}$ and $\mathscr{A}_h(u_h), f_h \in \mathbb{V}_h \subseteq \mathbb{V}$.

Suppose that (23.11) also has a unique solution. We can then evaluate a numerical approximation of $J(u)$ as $J_h(u_h)$, where J_h is an approximation of J, i.e.,

$$J_h(u_h) \approx J(u).$$

The problem we are interested in is the following: "Can we find a correction $\widehat{J_h(u_h)}$ for $J_h(u_h)$ such that

$$\left|\widehat{J_h(u_h)} - J(u)\right| < |J_h(u_h) - J(u)|, \tag{23.12}$$

and the cost for the correction is not more than the cost of solving (23.11) plus evaluating $J_h(u_h)$?"

Suppose J is Lipschitz continuous with Lipschitz constant c_J, then

$$\begin{aligned}|J_h(u_h) - J(u)| &\leq |J_h(u_h) - J(u_h)| + |J(u_h) - J(u)| \\ &\leq |J_h(u_h) - J(u_h)| + c_J \|u - u_h\|_{\mathbb{U}},\end{aligned}$$

which shows that the error in approximating the quantity of interest can be controlled by the discretization error for J_h (the first term on the right-hand side) and the solution error between the exact solution u and discretized solution u_h. For the simplicity of the exposition, we assume the error in discretizing J is the same order of the solution error. Thus, we can write

$$|J_h(u_h) - J(u)| \leq c \|u - u_h\|_{\mathbb{U}},$$

where c is some positive constant. That is,

$$|J_h(u_h) - J(u)| = \mathcal{O}(\|u - u_h\|_{\mathbb{U}}) \tag{23.13}$$

The estimate in (23.13) suggests two approaches to improve the approximation accuracy for the quantity of interest. In section 23.4, we derive an additive correction to $J_h(u_h)$ so that

$$\left|\widehat{J_h(u_h)} - J(u)\right| = \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right), \tag{23.14}$$

and thus yield a significantly smaller error: this is known as superconvergence [117]. In section 23.5, we present a technique to improve the approximate $J_h(u_h)$ by changing the discretization fidelity. This approach aims to intelligently reduce the error $\|u - u_h\|_{\mathbb{U}}$ while aiming to minimize the additional

cost in refining the discretization [15]. *At the heart of both approaches are the adjoint equation and its solution. We shall see that the adjoint solution can be used as either the key player in the correction term or a guideline on how to improve the discretization fidelity.* For both approaches, we deploy a similar optimization strategy to uncover the role of the adjoint: just like what did when deriving the Green function in subsection 13.2.2,

23.4 Error correction

The presentation of this section is inspired from the work in [117]. We, in subsection 23.4.1 however, derive the adjoint equation in a systematic manner and take into account additional errors (and thus remove some implicit assumptions in [117]). We also develop two different approaches in subsection 23.4.2 and subsection 23.4.3, the first of which is based on second-order Taylor expansion (23.4), and the second of which is based on the trapezoidal rule (23.6). As alluded from subsection 13.2.2, We consider the following artificial optimization problem

$$\min_{u\in\mathbb{U}} J(u), \qquad \text{subject to (23.10)}$$

which is trivial as we assume that (23.10) has a unique solution. Now applying the Lagrangian multiplier Theorem 9.3, the first order optimality condition (9.6) gives us the following adjoint equation

$$[\mathscr{D}\mathscr{A}(u)]^* v = -\mathscr{D}J(u), \tag{23.15}$$

where v is the adjoint solution (we shall assume that the adjoint equation (23.15) also has a unique solution).

Remark 23.2. Note that while the original equation (also known as the forward equation or the primal equation) (23.10) could be nonlinear in u, the adjoint equation (23.15) is always linear in the adjoint variable v.

In general, it may not be feasible to solve the adjoint equation (23.15) directly, and in that case we have to resort to some approximation. One approach is to discretize the adjoint equation with the same fidelity h to obtain

$$[\mathscr{D}\mathscr{A}_h(u_h)]^* v_h = -\mathscr{D}J_h(u_h). \tag{23.16}$$

There are at least two ways to understand the notations in (23.16), each of which will lead to a (slightly) different approach for improving the estimation error. The first view is to understand $\mathscr{D}\mathscr{A}_h(u_h)$ as first evaluating the Fréchet derivative $\mathscr{D}\mathscr{A}$ at u_h and then discretizing it with $\mathscr{D}\mathscr{A}_h$; similarly, we first evaluate $\mathscr{D}J$ at u_h and then discretize it to obtain $\mathscr{D}J_h(u_h)$. In practice, the order (e.g., evaluating the Fréchet derivative $\mathscr{D}\mathscr{A}$ at u_h and then discretizing

it with $\mathscr{D}\mathscr{A}_h$) is not important and in fact they are typically carried out at the same time. The second view is to evaluate $\mathscr{A}$ at u_h, then discretize $\mathscr{A}$ with $\mathscr{A}_h$, and then compute the Fréchet derivative of $\mathscr{A}_h(u_h)$ to obtain $\mathscr{D}\mathscr{A}_h(u_h)$; similarly, we evaluate J at u_h, then discretize J with J_h, and then compute the Fréchet derivative of $J_h(u_h)$ to obtain $\mathscr{D}J_h(u_h)$. It is straightforward to show (see Problem 23.3) that the second view can be obtained from the first order optimality condition of the following optimization problem

$$\min_{u_h} J_h(u_h), \quad \text{subject to (23.11)}.$$

It is important to point out that solving the discretized adjoint equation (23.16) is generally easier than solving the discrete forward equation (23.11), provided that $\mathscr{D}J_h(u_h)$ and $[\mathscr{D}\mathscr{A}_h(u_h)]^*$ can be evaluated[2], as the former is always linear in the adjoint solution v_h.

Till the rest of this section, we focus on the error correction using the first viewpoint of (23.16). We also assume that the disrete forward solution u_h in (23.11) and the discrete adjoint solution v_h in (23.16) are already computed.

23.4.1 First approach: first-order Taylor expansion

We begin by noticing that the adjoint equation, and hence the adjoint solution, involves the derivative of the quantity of interest $J(u)$ while our goal is to improve the approximation $J_h(u_h)$. *The natural approach to bring the two together is to invoke the Taylor expansion as it involves both function value and the derivative!* Indeed, applying (23.4) with J in place of $\mathcal{F}$, u_h in place of u, $u - u_h$ in place of v, and $n = 0$ gives

$$\begin{aligned} J(u) &= J(u_h) + \int_0^1 \mathscr{D}J(u_h + t(u - u_h); u - u_h)\, dt \\ &= J(u_h) + \left\langle \int_0^1 \mathscr{D}J(u_h + t(u - u_h))\, dt, u - u_h \right\rangle, \end{aligned} \tag{23.17}$$

where we have used the duality pairing form of Fréchet derivative along the direction $u - u_h$ and the linearity of the duality pairing (see section 9.2). Furthermore, we have assumed that we can switch between integral and the duality pairing. Let us denote

$$\overline{\mathscr{D}J(u_h, u)} := \int_0^1 \mathscr{D}J(u_h + t(u - u_h))\, dt.$$

[2] Note that there is no need to form the adjoint $[\mathscr{D}\mathscr{A}_h(u_h)]^*$ if one uses Krylov subspace methods. In that case, all we need is the ability to compute the action of the adjoint $[\mathscr{D}\mathscr{A}_h(u_h)]^*$ on an arbitrary function/vector.

A closer look reveals that $\overline{\mathscr{D}J\left(u_h,u\right)}$ is the average of the Fréchet derivative $\mathscr{D}J$ along the convex combination of u_h and u. As a result, it is symmetric in u_h and u (see Problem 23.4), i.e.,

$$\overline{\mathscr{D}J\left(u_h,u\right)}=\overline{\mathscr{D}J\left(u,u_h\right)}.$$

We thus arrive at

$$J\left(u\right)=J\left(u_h\right)+\left\langle\overline{\mathscr{D}J\left(u_h,u\right)},u-u_h\right\rangle. \tag{23.18}$$

Since the discretized equation (23.16) involves $\mathscr{D}J_h\left(u_h\right)$ instead of the average, we perform the standard adding and subtracting trick to obtain

$$J\left(u\right)=J\left(u_h\right)+\left\langle\mathscr{D}J_h\left(u_h\right),u-u_h\right\rangle+\left\langle\overline{\mathscr{D}J\left(u_h,u\right)}-\mathscr{D}J_h\left(u_h\right),u-u_h\right\rangle,$$

which, together with the discretized adjoint[3] equation (23.16), is equivalent to

$$J\left(u\right)=J\left(u_h\right)+\left\langle v_h,\mathscr{D}\mathscr{A}_h\left(u_h\right)\left(u_h-u\right)\right\rangle+\left\langle\overline{\mathscr{D}J\left(u_h,u\right)}-\mathscr{D}J_h\left(u_h\right),u-u_h\right\rangle, \tag{23.19}$$

where we have also invoked the adjoint definition in the second term of the right-hand side. Continuing the adding and subtracting trick gives

$$\begin{aligned}J\left(u\right)&=J\left(u_h\right)\\&\quad+\left\langle v_h,\mathscr{D}\mathscr{A}\left(u_h\right)\left(u_h-u\right)\right\rangle+\underbrace{\left\langle v_h,\left(\mathscr{D}\mathscr{A}_h\left(u_h\right)-\mathscr{D}\mathscr{A}\left(u_h\right)\right)\left(u_h-u\right)\right\rangle}_{\text{term I}}\\&\quad+\underbrace{\left\langle\overline{\mathscr{D}J\left(u_h,u\right)}-\mathscr{D}J\left(u_h\right),u-u_h\right\rangle}_{\text{term II}}+\underbrace{\left\langle\mathscr{D}J\left(u_h\right)-\mathscr{D}J_h\left(u_h\right),u-u_h\right\rangle}_{\text{term III}}.\end{aligned}$$

Suppose the errors in discretizing $\mathscr{D}\mathscr{A}$ and $\mathscr{D}J$ are the same order of the forward solution error $\left\|u-u_h\right\|_{\mathbb{U}}$, then term I and term III are high-order as they scale like $\mathcal{O}\left(\left\|u-u_h\right\|_{\mathbb{U}}^2\right)$. Term II is also high-order, as we know show. We have

$$\left|\text{term II}\right|=\left|\int_0^1\left\langle\left[\mathscr{D}J\left(u_h+t\left(u-u_h\right)\right)-\mathscr{D}J\left(u_h\right)\right],u-u_h\right\rangle\,dt\right|.$$

Either applying the Taylor expansion (23.4) for $\mathscr{D}J\left(u_h+t\left(u-u_h\right)\right)$ around $\hat{u}$ and assume that the second order Fréchet derivative of J is bounded (see Problem 23.5), or assuming that $\mathscr{D}J$ is Lipschitz continuous[4], i.e.,

[3] Recall the discussion in Remark 23.1 that we assume the discrete adjoint equation (23.16) is an identity in $\mathbb{U}$ and the adjoint $\left[\mathscr{D}\mathscr{A}_h\left(u_h\right)\right]^*$ is defined as a linear map from $\mathbb{V}$ to $\mathbb{U}$. Otherwise, we could not substitute (23.16), had it been valid only in a subspace $\mathbb{U}_h\in\mathbb{U}$.

[4] The two approaches are essentially equivalent.

$$|[\mathscr{D}J(u_h + t(u - u_h)) - \mathscr{D}J(u_h)]| = \mathcal{O}(t\|u - u_h\|_{\mathbb{U}})$$

we have

$$|\text{term II}| = \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right).$$

Therefore, we have shown

$$J(u) = J(u_h) + \underbrace{\langle v_h, \mathscr{D}\mathscr{A}(u_h)(u_h - u)\rangle}_{\text{term IV}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right). \tag{23.20}$$

Next, applying the Taylor expansion (23.4) with $n = 1$ for $\langle v_h, \mathscr{A}(u)\rangle_{\mathbb{V}}$ around u_h, we have

$$\begin{aligned}\langle v_h, \mathscr{A}(u)\rangle_{\mathbb{V}} = \langle v_h, \mathscr{A}(u_h)\rangle_{\mathbb{V}} - \underbrace{\langle v_h, \mathscr{D}\mathscr{A}(u_h)(u_h - u)\rangle}_{\text{term IV}} \\ + \int_0^1 \left\langle v_h, \mathscr{D}^2\mathscr{A}(u_h + t(u - u_h); u - u_h, u - u_h)\right\rangle (1 - t)\, dt.\end{aligned}$$

Suppose that $\mathscr{D}^2\mathscr{A}(u_h + t(u - u_h))$ is bounded, then

$$\int_0^1 \left\langle v_h, \mathscr{D}^2\mathscr{A}(u_h + t(u - u_h); u - u_h, u - u_h)\right\rangle (1 - t)\, dt = \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

from which it follows that

$$\begin{aligned}\text{term IV} &= \langle v_h, \mathscr{A}(u_h) - \mathscr{A}(u)\rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right) \\ &= \langle v_h, \mathscr{A}_h(u_h) - \mathscr{A}(u)\rangle_{\mathbb{V}} + \langle v_h, \mathscr{A}(u_h) - \mathscr{A}_h(u_h)\rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right) \\ &= \langle v_h, f_h - f\rangle_{\mathbb{V}} + \langle v_h, \mathscr{A}(u_h) - \mathscr{A}_h(u_h)\rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),\end{aligned}$$

where we have used (23.10) and (23.11) in the first term of the second equality to obtain the last equality.

We conclude that

$$\begin{aligned}J(u) &= J(u_h) + \langle v_h, f_h - f\rangle_{\mathbb{V}} + \langle v_h, \mathscr{A}(u_h) - \mathscr{A}_h(u_h)\rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right) \\ &= J_h(u_h) + \langle v_h, f_h - f\rangle_{\mathbb{V}} + \underbrace{J(u_h) - J_h(u_h)}_{\text{term A}} \\ &\quad + \underbrace{\langle v_h, \mathscr{A}(u_h) - \mathscr{A}_h(u_h)\rangle_{\mathbb{V}}}_{\text{term B}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),\end{aligned}$$

which gives us three scenarios:

i Suppose that there is no discretization error in J and $\mathscr{A}$, that is,

$$J(u_h) = J_h(u_h) \text{ and } \mathscr{A}(u_h) = \mathscr{A}_h(u_h).$$

Then define

$$\widehat{J_h(u_h)} := J_h(u_h) + \langle v_h, f_h - f\rangle_{\mathbb{V}},$$

we trivially obtain the superconvergence result (23.14).

ii) Suppose that we can evaluate $J(u_h)$ and $\mathscr{A}(u_h)$ exactly. We then define a computable corrected quantity of interest as

$$\widehat{J_h(u_h)} := J_h(u_h) + \langle v_h, f_h - f\rangle_{\mathbb{V}} + J(u_h) - J_h(u_h) + \langle v_h, \mathscr{A}(u_h) - \mathscr{A}_h(u_h)\rangle_{\mathbb{V}}, \quad (23.21)$$

and this immediately yields the desirable estimate (23.14).

iii) If we cannot evaluate $J(u_h)$ and $\mathscr{A}(u_h)$ exactly, we can still achieve (23.14) if the discretization for J and $\mathscr{A}$ is an order of magnitude smaller than the approximation for u. In particular, assume that we design our numerical discretization for J and $\mathscr{A}$ such that

$$|J(u_h) - J_h(u_h)| = \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$
$$\|\mathscr{A}(u_h) - \mathscr{A}_h(u_h)\|_{\mathbb{V}} = \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right).$$

Then, again, defining $\widehat{J_h(u_h)}$ as in (23.21) gives the desirable estimate (23.14).

23.4.2 Second approach: second-order Taylor expansion

The derivation in subsection 23.4.1 is based on the zero-order Taylor expansion of $J(u)$ (see (23.17)). We then perform the add-add-subtract trick (which could be unintuitive) to make the derivative of J appear in order to invoke the discrete adjoint equation (23.16). In this section, we provide a succinct and elegant approach using the first-order Taylor expansion (23.4) for $J(u)$. In particular, we have

$$J(u) = J(u_h) + \langle \mathscr{D}J(u_h), u - u_h\rangle + \int_0^1 \mathscr{D}J^2(u_h + t(u - u_h); u - u_h, u - u_h)(1 - t)\, dt. \quad (23.22)$$

At this point, we could have used the discrete adjoint equation (23.16) if the second term on the right-hand side were with $\mathscr{D}J_h$. But this is simple as we can add and subtract to have

$$J(u) = J(u_h) + \langle \mathscr{D}J_h(u_h), u - u_h \rangle + \langle \mathscr{D}J(u_h) - \mathscr{D}J_h(u_h), u - u_h \rangle + \int_0^1 \mathscr{D}J^2(u_h + t(u - u_h); u - u_h, u - u_h)(1 - t)\, dt.$$

Now we can invoke the discrete adjoint equation (23.16) to arrive at

$$J(u) = J(u_h) + \langle v_h, \mathscr{D}\mathscr{A}_h(u_h)(u_h - u) \rangle + \underbrace{\langle \mathscr{D}J(u_h) - \mathscr{D}J_h(u_h), u - u_h \rangle}_{\text{term A}} + \underbrace{\int_0^1 \mathscr{D}J^2(u_h + t(u - u_h); u - u_h, u - u_h)(1 - t)\, dt}_{\text{term B}}.$$

Since term A is the same as term III in subsection 23.4.1, it scales as $\mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right)$. For term B, similar to subsection 23.4.1, if we assume that the second-order Fréchet derivative $\mathscr{D}J^2(\cdot)$, as a linear operator from $\mathbb{U} \times \mathbb{U} \to \mathbb{R}$ is (essentially) bounded[5] along the line connecting u_h and u, then we have

$$J(u) = J(u_h) + \langle v_h, \mathscr{D}\mathscr{A}_h(u_h)(u_h - u) \rangle + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

which is the same as (23.20). Thus, we can follow the rest of the subsection 23.4.1, including how to construct an improved quantity of interest $\widehat{J_h(u_h)}$.

23.4.3 Third approach: Trapezoidal trick

In the third approach, similar to the first approach in subsection 23.4.1, the starting point is the zero-order Taylor expansion for $J(u)$:

$$J(u) = J(u_h) + \int_0^1 \mathscr{D}J(u_h + t(u - u_h); u - u_h)\, dt. \tag{23.23}$$

The key is then to invoke the trapezoidal rule (23.6) for the second term to have

[5] Note that it is sufficient to assume that the $\mathscr{D}J(\cdot)$ is Lipschitz continuous along the line connecting u_h and u

$$\int_0^1 \mathscr{D}J\left(u_h + t\left(u - u_h\right); u - u_h\right)\, dt$$
$$= \frac{1}{2}\left\langle \mathscr{D}J\left(u_h\right), u - u_h\right\rangle + \frac{1}{2}\left\langle \mathscr{D}J\left(u\right), u - u_h\right\rangle$$
$$- \int_0^1 \left(t - \frac{1}{2}\right) \mathscr{D}^2 J\left(u_h + t\left(u - u_h\right); u - u_h, u - u_h\right)\, dt,$$

which, after applying Taylor expansion for the second term on the right-hand side, becomes

$$\int_0^1 \mathscr{D}J\left(u_h + t\left(u - u_h\right); u - u_h\right)\, dt$$
$$= \left\langle \mathscr{D}J\left(u_h\right), u - u_h\right\rangle + \int_0^1 \mathscr{D}J^2\left(u_h + t\left(u - u_h\right); u - u_h; u - u_h\right)\left(1 - t\right)\, dt.$$

Thus,

$$J\left(u\right) = J\left(u_h\right) + \left\langle \mathscr{D}J\left(u_h\right), u - u_h\right\rangle$$
$$+ \int_0^1 \mathscr{D}J^2\left(u_h + t\left(u - u_h\right); u - u_h, u - u_h\right)\left(1 - t\right)\, dt,$$

which is exactly the starting point of subsection 23.4.2. Thus, we have shown that starting from the first approach with judicious use of the trapezoidal rule and Taylor expansion, we recover the second approach. In other words, the first and second approaches are equivalent.

23.5 A posterior error estimation

23.5.1 A variational formulation and its discretization

The setting in this section, in a sense, is more general than the one in section 23.4. This is facilitated by a variational setting. Equipped with Fréchet derivative definition via duality pairing (see section 9.2 of Chapter 9) and the adjoint definition with duality pairing (see the third footnote of Chapter 13), we can consider a more general setting without additional effort or difficulty. We start by considering

$$\mathscr{A} : \mathbb{U} \ni u \mapsto \mathscr{A}\left(u\right) \in \mathbb{V}^* \text{ and } f \in \mathbb{V}^*,$$

and writing (23.10) in an equivalent variational form with duality pairing in $\mathbb{V}$:

$$a\left(u; z\right) := \left\langle \mathscr{A}\left(u\right), z\right\rangle_{\mathbb{V}} = \left\langle f, z\right\rangle_{\mathbb{V}} =: \ell\left(z\right), \quad \forall z \in \mathbb{V}. \tag{23.24}$$

When $\mathscr{A}$ is a linear operator, variational settings and their well-posedness can be referred to Chapter 15.

Remark 23.3. In the finite element literature, $\mathbb{U}$ and $\mathbb{V}$ are commonly called the trial and test spaces.

Again, our convention is that $a(\cdot;\cdot)$ depends linearly on any argument after the semicolon. Let $\mathbb{U}_h \subset \mathbb{U}$, $\mathbb{V}_h \subset \mathbb{V}$, and consider the following abstract discretization[6] of (23.24):

$$a(u_h; z_h) := \langle \mathscr{A}(u_h), z_h \rangle_{\mathbb{V}} = \ell(z_h), \quad \forall z_h \in \mathbb{V}_h. \tag{23.25}$$

The "Galerkin" orthogonality

$$\langle \mathscr{A}(u_h), z_h \rangle_{\mathbb{V}} - \langle \mathscr{A}(u), z_h \rangle_{\mathbb{V}} = 0, \quad \forall z_h \in \mathbb{V}_h, \tag{23.26}$$

is an immediate consequence of (23.24) and (23.25).

Remark 23.4. Compared to (23.11) (see Remark 23.1), the discretized forward equation (23.25) is the restriction of the forward equation (23.24) in subspaces $\mathbb{U}_h \times \mathbb{V}_h$. In particular, we look for a solution $u_h \in \mathbb{U}_h \subset \mathbb{U}$ for all test functions $z_h \in \mathbb{V}_h \subset \mathbb{V}$. We would like to point out that the discretized variational equation (23.25) is typically a result of a finite element method in which we do not discretize the operator $\mathscr{A}$, but analytically evaluate it on the approximate solution u_h (which are typically piecewise polynomials). In practice, most of the variational settings have $\mathbb{U}$ and $\mathbb{V}$ as Sobolev spaces with predefined inner-product products (see subsection 13.1.5). In these cases, the duality pairings in (23.25) are in fact inner products and can be evaluated in a piecewise fashion. We assume that both sides of (23.25) can be evaluated exactly though they are typically approximated with some quadrature rules. We shall make the same assumption for the discrete adjoint equation (23.28) in the following. Taking into account the quadrature error is possible, but is tedious/cumbersome and yet does not add much value to our exposition.

Similarly, let us generalize the adjoint equation (23.15) using duality pairing in $\mathbb{U}$ as

$$\left\langle [\mathscr{D}\mathscr{A}(u)]^* v, w \right\rangle_{\mathbb{U}} = -\langle \mathscr{D}J(u), w \rangle_{\mathbb{U}}, \quad \forall w \in \mathbb{U}, \tag{23.27}$$

and its discretization in $\mathbb{V}_h$ and $\mathbb{U}_h$ as

$$\left\langle [\mathscr{D}\mathscr{A}(u_h)]^* v_h, w_h \right\rangle_{\mathbb{U}} = -\langle \mathscr{D}J(u_h), w_h \rangle_{\mathbb{U}}, \quad \forall w_h \in \mathbb{U}_h, \tag{23.28}$$

where, similar to the discretization strategy for the forward equation, we restrict the adjoint equation (23.27) on the subspaces $\mathbb{U}_h \times \mathbb{V}_h$ to obtain the discretized adjoint equation (23.28).

[6] Note that in practice, we may have to discretize the duality pairing and/or inner products as well, but we ignore this for the clarity of the exposition.

23.5.2 *From error correction to* a posteriori *error estimation*

In this section, we first carry out a similar procedure that we did in section 23.4 for the variational approach. The Taylor expansion (23.18) still holds. Guided by subsection 23.4.2, we start with (23.22):

$$J(u) = J(u_h) + \langle \mathscr{D}J(u_h), u - u_h \rangle_{\mathbb{U}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right).$$

The problem is that we can not use the left-hand side of the discretized adjoint equation (23.28) since it is only valid in the subspace $\mathbb{U}_h$, while $u \notin \mathbb{U}_h$ in general. As a result, we cannot substitute (23.28) into the second term of $\langle \mathscr{D}J(u_h), u - u_h \rangle$ as we have done in (23.19). Thus, the correct path should be through the (continuous) adjoint (23.27), but we need $\mathscr{D}J(u)$ and this can be achieve by add-and-subtract trick to have

$$\begin{aligned}\langle \mathscr{D}J(u_h), u - u_h \rangle_{\mathbb{U}} &= \langle \mathscr{D}J(u), u - u_h \rangle_{\mathbb{U}} + \langle \mathscr{D}J(u_h) - \mathscr{D}J(u), u - u_h \rangle_{\mathbb{U}} \\ &\underbrace{=}_{(23.27)} \langle v, \mathscr{D}\mathscr{A}(u)(u - u_h) \rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right).\end{aligned}$$

Thus,

$$J(u) = J(u_h) + \langle v, \mathscr{D}\mathscr{A}(u)(u - u_h) \rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right), \tag{23.29}$$

which is almost the same as (23.19) except we have v instead of v_h in the second term on the right-hand side. Again, with the add-and-subtract trick we have

$$\langle v, \mathscr{D}\mathscr{A}(u)(u - u_h) \rangle_{\mathbb{V}} = \langle v_h, \mathscr{D}\mathscr{A}(u)(u - u_h) \rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

where we have assumed that the error in the adjoint discrete solution $\|v - v_h\|_{\mathbb{V}}$ is the same order of the error in the forward solution $\|u - u_h\|_{\mathbb{V}}$, and that $\mathscr{D}\mathscr{A}(u)$ is bounded.

Now, similar to manipulating (23.20), using Taylor expansion (23.4) with $n = 1$ for $\langle v_h, \mathscr{A}(u_h) \rangle_{\mathbb{V}}$ around u, yields

$$\langle v_h, \mathscr{D}\mathscr{A}(u)(u - u_h) \rangle_{\mathbb{V}} = \underbrace{\langle v_h, \mathscr{A}(u_h) \rangle_{\mathbb{V}} - \langle v_h, \mathscr{A}(u) \rangle_{\mathbb{V}}}_{=0 \text{ due to the Galerin orthogonality (23.26)}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

where we have assumed that $\mathscr{D}^2\mathscr{A}(u_h + t(u - u_h))$ is bounded.[7]

We conclude that

[7] Note that correct term is $\mathscr{D}^2\mathscr{A}(u + t(u_h - u))$, but Problem 23.4 tells us that they are the same.

$$J(u) = J(u_h) + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

but this is not a useful estimate as it does suggest a way to improve $J(u_h)$!

To avoid the Galerkin orthogonality we go back to (23.29) and Taylor expansion (23.4) with $n = 1$ for $\langle v, \mathscr{A}(u_h)\rangle_{\mathbb{V}}$ around u to arrive at

$$J(u) = J(u_h) + \langle v, \mathscr{A}(u_h) - \mathscr{A}(u)\rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

which, after inserting the Galerkin orthogonality (23.26), becomes

$$J(u) = J(u_h) + \langle v - v_h, \mathscr{A}(u_h) - \mathscr{A}(u)\rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

which is equivalent to

$$J(u) = J(u_h) + \langle v - v_h, \mathscr{A}(u_h) - f\rangle_{\mathbb{V}} + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right),$$

where we have used (23.24). Now if we define the residual of the forward equation (23.24) at u_h projected on $v - v_h$ as

$$\mathcal{R}(u_h; v - v_h) := \langle \mathscr{A}(u_h), v - v_h\rangle_{\mathbb{V}} - \langle f, v - v_h\rangle_{\mathbb{V}} = \langle \mathscr{A}(u_h) - f, v - v_h\rangle_{\mathbb{V}},$$

we arrive at the weighted residual form [56, 15, 5]

$$J(u) = J(u_h) + \mathcal{R}(u_h; v - v_h) + \mathcal{O}\left(\|u - u_h\|_{\mathbb{U}}^2\right). \tag{23.30}$$

The problem is that we do not know the exact adjoint solution v, and thus the weighted residual, if used as a correction term, is generally not available.

The fact that the weighted residual $\mathcal{R}(u_h; v - v_h)$ is not computable does not prevent us from approximating it. We can then explore the residual approximation to refine the discretization to improve the accuracy of $J(u_h)$. To that end, suppose the domain $\Omega \subset \mathbb{R}^n$ of interest (on which we consider all the quantities including the forward and adjoint solutions, and their discrete approximations) is open and bounded. We also write $\mathbb{U}(\Omega)$ and $\mathbb{V}(\Omega)$ to highlight the fact that both trial and test space depend on the domain Ω. Let us decompose Ω into N non-overlapping subdomains $\mathsf{D}_k, k = 1, \ldots, \mathsf{N}$ such that

$$\overline{\Omega} = \cup_{k=1}^{\mathsf{N}} \overline{\mathsf{D}_k},$$

where the overline denotes the closure of a set (see Definition 5.8). We further assume that the weighted residual can be decomposed[8] as

[8] Note that this is not a significant limitation, as $\mathbb{V}$ is typically an $\mathbb{L}^2$-based Sobolev space (see Definition 13.15) in most of applications.

$$\mathcal{R}\left(u_h; v - v_h\right) = \langle \mathscr{A}\left(u_h\right) - f, v - v_h \rangle_{\mathbb{V}(\Omega)} = \sum_{k=1}^{\mathsf{N}} \underbrace{\langle \mathscr{A}\left(u_h\right) - f, v - v_h \rangle_{\mathbb{V}(\mathsf{D}_k)}}_{=:\mathcal{R}_k(u_h; v - v_h)}$$

where the local weighted residual $\mathcal{R}_k\left(u_h; v - v_h\right)$ is evaluated on each subdomain D_k. By invoking the Cauchy-Schwarz inequality we arrive at

$$\left|\mathcal{R}\left(u_h; v - v_h\right)\right| \leq \sqrt{\sum_{k=1}^{\mathsf{N}} \left\|\mathscr{A}\left(u_h\right) - f\right\|^2_{\mathbb{V}(\mathsf{D}_k)}} \sqrt{\sum_{k=1}^{\mathsf{N}} \left\|v - v_h\right\|^2_{\mathbb{V}(\mathsf{D}_k)}}.$$

Since we assume that the discrete forward and adjoint solutions u_h and v_h are already computed, the element residual $\left\|\mathscr{A}\left(u_h\right) - f\right\|_{\mathbb{V}(\mathsf{D}_k)}$ is computable. The adjoint error $\left\|v - v_h\right\|^2_{\mathbb{V}(\mathsf{D}_k)}$ can be estimated if we have a more accurate approximation of v than v_h. Two popular approaches are:

1. Solve the discrete adjoint problem (23.28) with higher fidelity (e.g., refining the mesh or increasing the solution order for finite element methods) to obtain a more accurate discrete adjoint solution ${v_h}^f$, and then replace $\left\|v - v_h\right\|^2_{\mathbb{V}(\mathsf{D}_k)}$ with $\left\|{v_h}^f - v_h\right\|^2_{\mathbb{V}(\mathsf{D}_k)}$. For nonlinear forward problem (23.25), solving for u_h is significantly more costly than solving for the discrete adjoint problem (23.28) as the latter is always linear. For such a case, solving for ${v_h}^f$ does not incur more cost than solving the discrete forward problem (23.25). If the forward problem is linear forward problem (23.25) or when we do not want to solve an additional adjoint equation to estimate the adjoint error, we can resort to the second approach.
2. Post-process the discrete adjoint solution v_h (such as high-order interpolation for ${v_h}^f$ in an element D_k based on v_h from the neighboring element) to obtain a more accurate representation ${v_h}^f$ (see, e.g., [35, 4, 144]). This approach is an efficient approach as it can be carried out for all elements in a completely parallel fashion. However, the error estimation may not be the same accuracy/quality as the first approach as the adjoint solution needs to be sufficiently smooth for the interpolation to be a more accurate representation.

Problems

Problem 23.1. Provide a derivation/proof for (23.3).

Problem 23.2. Using (23.4) to derive the error for the trapezoidal rule

Problem 23.3. Show that the first-order optimality condition of

$$\min_{u_h} J_h\left(u_h\right), \quad \text{subject to (23.11)}$$

is the second view of (23.16).

Problem 23.4. Prove the following symmetry

$$\overline{\mathscr{D}J\left(u_h,u\right)}=\overline{\mathscr{D}J\left(u,u_h\right)}.$$

Problem 23.5. Use the Taylor expansion (23.4) to show that

$$\left\langle\left[\mathscr{D}J\left(u_h+t\left(u-u_h\right)\right)-\mathscr{D}J\left(u_h\right)\right],u-u_h\right\rangle=\mathcal{O}\left(\left\|u-u_h\right\|_{\mathbb{U}}^2\right),$$

under some sufficient condition on the second order Fréchet derivative of J.

Chapter 24
Reproducing Kernel Hilbert Spaces

Abstract Due to the ubiquitous nature of Reproducing Kernel Hilbert Spaces (RKHS) in mathematics and recently in machine learning, this chapter exhibits the important role of the adjoint in the theory of RKHS. We shall see that the Mercer theorem is a direct consequence of the Hilbert-Schmidt Theorem 14.1. We present RKHS from two dual perspectives: i) starting from a RKHS and deriving the associated kernel, and ii) starting from a kernel and deriving the RKHS. Due to the one-to-one association of a RKHS and its kernel (which we will also prove), the two perspectives are equivalent. Besides our own findings and derivations, the majority of the materials in this chapter can be found in [62, 98, 114, 40] and the references therein. The prerequisites for this chapter are:

- Chapter 5, Chapter 13 and Chapter 14
- ...

24.1 From a Reproducing Kernel Hilbert Space to its kernel

Definition 24.1 (Reproducing Kernel Hilbert Space). Let X be a set. A vector space $\mathscr{H}$ on X with value in $\mathbb{F}$ (either real or complex) is a Reproducing Kernel Hilbert Space (RKHS) if

H1) $\mathscr{H}$ is a Hilbert space with an inner product $(\cdot,\cdot)_{\mathscr{H}}$, and
H2) the pointwise evaluation functional $\mathrm{e}_x : \mathscr{H} \ni f \mapsto \mathrm{e}_x(f) := f(x) \in \mathbb{F}$ is bounded for any $x \in X$.

From **H2)**, a RKHS is special in the sense that, unlike a general Hilbert space (such as Sobolev spaces in Chapter 13), pointwise value/evaluation for functions in a RKHS is well-defined. Moreover, e_x is a linear and continuous functional on $\mathscr{H}$, and thus resides in $\mathscr{H}^*$. By the Riesz representation

Theorem 5.1, there exists a unique $\mathsf{K}_x \in \mathscr{H}$ such that

$$\|e_x\|_{\mathscr{H}^*} = \|\mathsf{K}_x\|_{\mathscr{H}},$$

and

$$f(x) = \mathrm{e}_x(f) = (\mathsf{K}_x, f)_{\mathscr{H}}, \tag{24.1}$$

which is known as the reproducing property. *We can see that K_x is a generalized Green function (see subsection 13.2.1.1 for a detailed discussion), and this will also be clear in Example 24.4 for a specific example. Furthermore, e_x is simply the Dirac delta distribution (see Definition 13.11) if the test space $\mathcal{D}(X)$ is dense in $\mathscr{H}$.*

Definition 24.2 (The induced kernel from a RKHS $\mathscr{H}$). The kernel $K(\cdot,\cdot) : X \times X \to \mathbb{F}$ associated with a RKHS $\mathscr{H}$ is defined as

$$\mathsf{K}(x,y) := \mathsf{K}_y(x), \quad \forall x, y \in X.$$

The induced kernel $\mathsf{K}(x,y)$ makes sense since, from (24.1), for any $y \in X$, $\mathsf{K}_y \in \mathscr{H}$ and thus the pointwise evaluation $\mathsf{K}_y(x)$ is meaningful for any $x \in X$. Furthermore, also from the reproducing property (24.1), we have

$$\mathsf{K}(x,y) := \mathsf{K}_y(x) = (\mathsf{K}_x, \mathsf{K}_y)_{\mathscr{H}}.$$

Thus, from the natural inner product (5.5) in the dual space $\mathscr{H}^*$, the two-point kernel $\mathsf{K}(x,y)$ is exactly the inner product of the two pointwise evaluation functionals e_x and e_y (see (5.5)):

$$\mathsf{K}(x,y) = (\mathsf{K}_x, \mathsf{K}_y)_{\mathscr{H}} = (\mathrm{e}_y, \mathrm{e}_x)_{\mathscr{H}^*}.$$

Two important observations from the induced kernel in Definition 24.2 are in order.

K1) $\mathsf{K}(\cdot,\cdot)$ is symmetric as

$$\mathsf{K}(x,y) = (\mathsf{K}_x, \mathsf{K}_y)_{\mathscr{H}} = \overline{(\mathsf{K}_y, \mathsf{K}_x)_{\mathscr{H}}} = \overline{\mathsf{K}(y,x)},$$

where we have used the symmetry of the inner product in $\mathscr{H}$ (see Chapter 4). As a result, we have

$$\|e_x\|^2_{\mathscr{H}^*} = \|\mathsf{K}_x\|^2_{\mathscr{H}} = (\mathsf{K}_x, \mathsf{K}_x)_{\mathscr{H}} = \mathsf{K}(x,x). \tag{24.2}$$

K2) $\mathsf{K}(\cdot,\cdot)$ is symmetric positive definite,[1] as for any n number of points $\{x^i\}_{i=1}^n$, the corresponding matrix $\boldsymbol{K} \in \mathbb{F}^{n\times n}$ with $\boldsymbol{K}_{ij} = \mathsf{K}(x^i, x^j)$ is symmetric semi-positive definite:

[1] It is in fact semi-positive definite, but for historical reason (see, e.g., [114]) we shall use positive definite.

$$\overline{\boldsymbol{\alpha}}^T \boldsymbol{K} \boldsymbol{\alpha} = \sum_{i,j=1}^{n} \overline{\boldsymbol{\alpha}_i} \mathsf{K}\left(x^i, x^j\right) \boldsymbol{\alpha}_j = \left(\sum_{i=1}^{n} \boldsymbol{\alpha}_i \mathsf{K}_{x^i}, \sum_{j=1}^{n} \boldsymbol{\alpha}_j \mathsf{K}_{x^j} \right)_{\mathscr{H}} \geq 0, \forall \boldsymbol{\alpha} \in \mathbb{F}^n.$$

Example 24.1 ($\ell^2(\mathbb{N})$ is a RKHS). First, let us consider the special finite dimensional Hilbert space $\mathbb{C}^n$ with the standard inner product $(\boldsymbol{f}, \boldsymbol{g})_{\mathbb{C}^n} := \sum_{i=1}^{n} \overline{\boldsymbol{f}_i} \boldsymbol{g}_i$. If we set $X = \{1, \ldots, n\}$, then any $x = i \in X$ each vector $\boldsymbol{f} \in \mathbb{C}^n$ is a function on X with values in $\mathbb{C}$, i.e., $|\boldsymbol{f}(x)| = |\boldsymbol{f}_i| \leq \|\boldsymbol{f}\|_{\mathbb{C}^n}$. Thus, each pointwise linear evaluation functional e_i is continuous, and returns the ith component of a vector in $\mathbb{C}^n$. Its unique Riesz representation K_i is simply the ith canonical basis vector of $\mathbb{C}^n$. The two-point kernel in this case reads

$$\mathsf{K}(i,j) = (\mathsf{K}_i, \mathsf{K}_j)_{\mathbb{C}^n} = \delta_{ij} = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{otherwise.} \end{cases}$$

As a result, $\|e_i\|_{[\mathbb{C}^n]^*} = \|\mathsf{K}_i\|_{\mathbb{C}^n} = \sqrt{\mathsf{K}(i,i)} = 1$. Moreover, the associated matrix $\boldsymbol{K}$ is simply the identity matrix.

Now recall that $\ell^2(\mathbb{N}) := \left\{ \boldsymbol{f} \in \mathbb{R}^\infty : \sum_{n=1}^{\infty} |\boldsymbol{f}_n|^2 < \infty \right\}$ with the standard inner product

$$(\boldsymbol{f}, \boldsymbol{g})_{\ell^2(\mathbb{N})} := \sum_{n=1}^{\infty} \overline{\boldsymbol{f}_n} \boldsymbol{g}_n, \quad \forall \boldsymbol{f}, \boldsymbol{g} \in \ell^2(\mathbb{N}),$$

and the induced norm $\|\boldsymbol{f}\|_{\ell^2(\mathbb{N})} = \sqrt{(\boldsymbol{f}, \boldsymbol{f})_{\ell^2(\mathbb{N})}}$. If we set $X = \mathbb{N}$, then clearly $\boldsymbol{f}(x)$ returns the mth coordinate of $\boldsymbol{f}$ for any $x = m \in \mathbb{N}$ and $\boldsymbol{f} \in \ell^2(\mathbb{N})$, that is, $|e_m(\boldsymbol{f})| = |\boldsymbol{f}_m| \leq \|\boldsymbol{f}\|_{\ell^2(\mathbb{N})}$. It follows that e_m is linear and continuous, and its unique Riesz representation K_m is exactly the mth canonical basis vector[2] of $\ell^2(\mathbb{N})$. The two-point kernel is again $\mathsf{K}(n,m) = \delta_{nm}$, and the associated matrix $\boldsymbol{K}$ for any finite number of points is the identity matrix.

Example 24.2 ($\mathbb{L}^2(X)$, $X \subseteq \mathbb{R}$ is not a RKHS on X). The $\mathbb{L}^2(\Omega)$ is defined in Definition 13.7. It would be natural to think that $\mathbb{L}^2(X)$ is the continuous version of $\ell^2(\mathbb{N})$, and thus $\mathbb{L}^2(\Omega)$ could be a RKHS. This is however not true and we now provide a counterexample. Consider $X = [0,1]$ and a sequence of functions

$$f_n(x) := \begin{cases} \sqrt{n} & \text{if } 0 \leq x \leq 1/n \\ 0 & \text{otherwise.} \end{cases}$$

It is clear that $\|f_n\|_{\mathbb{L}^2(X)} = 1$ for all $n \in \mathbb{N}$. By definition, the operator norm of the evaluation functional at $x = 0$, suppose it exists, is given by

[2] $\mathsf{K}_m = (0, \ldots, 0, 1, 0, \ldots)$, where 1 is at the mth location.

$$\|e_0\|_{[\mathbb{L}^2(X)]^*} := \sup_{f\in\mathbb{L}^2(X)} \frac{|f(0)|}{\|f\|_{\mathbb{L}^2(X)}} \geq |f_n(0)| = \sqrt{n} \xrightarrow{n\to\infty} \infty,$$

which is a contradiction. Thus $\mathbb{L}^2(X)$ for $X \subseteq \mathbb{R}^n$ is not a RKHS, as **H2)** does not hold.

Remark 24.1. The counterexample in Example 24.2 exposes the intrinsic difference between $\ell^2(\mathbb{N})$ and $\mathbb{L}^2(X)$ in that $X \subseteq \mathbb{R}$ is a continuum with Lebesgue measure while $\mathbb{N}$ is discrete with counting measure. In particular, we can allow the function $f_n(x)$ to grow unboundedly in vanishing small interval $[0, 1/n]$ and yet $f_n \in \mathbb{L}^2(X)$. As a result, we can squeeze in an unbounded amount of value/mass $\sqrt{n}$ at the point 0 for a function in $Ls^2(X)$. This is not possible for $\ell^2(\mathbb{N})$.

Example 24.3 ($\left[\mathbb{L}^2(\Omega)\right]^*$, $\Omega \subseteq \mathbb{R}^n$ is a RKHS on $\mathbb{L}^2(\Omega)$). As we have seen from Example 24.2 is not a RKHS. However, its topological dual $\left[\mathbb{L}^2(\Omega)\right]^*$ with the canonical inner product (see (5.5))

$$(\varphi, \eta)_{[\mathbb{L}^2(\Omega)]^*} = \overline{(\mathcal{R}^*\varphi, \mathcal{R}^*\eta)_{\mathbb{L}^2(\Omega)}},$$

is a RKHS on $\mathbb{L}^2(\Omega)$. Indeed, let us denote $X = \mathbb{L}^2(\Omega)$ and $\mathscr{H} = \left[\mathbb{L}^2(\Omega)\right]^*$. From Problem 5.13 we know that $\mathscr{H}$ is Hilbert space. Furthermore, each pointwise evaluation is given as

$$|e_x(\varphi)| = |\varphi(x)| \leq \|\varphi\|_{\mathscr{H}} \|x\|_X, \quad \forall x \in X,$$

which is linear and bounded. Since

$$(\mathsf{K}_x, \varphi)_{\mathscr{H}} = e_x(\varphi) = \varphi(x) = (\mathcal{R}^*\varphi, x)_X = (\varphi, \mathcal{R}x)_{\mathscr{H}},$$

the induced kernel is given by

$$\mathsf{K}(x, y) = (\mathsf{K}_y, \mathsf{K}_x)_{\mathscr{H}} = (\mathsf{K}_x, \mathcal{R}y)_{\mathscr{H}} = (\mathcal{R}y, \mathcal{R}x)_{\mathscr{H}} = \overline{(y, x)_X} = (x, y)_X.$$

That is, the two-point kernel is nothing more than the inner product of the corresponding two points in X.

Remark 24.2. As discussed in Remark 5.4, we may not need to distinguish $\mathbb{L}^2(\Omega)$ and its topological dual $\left[\mathbb{L}^2(\Omega)\right]^*$ as they are isometric to each other and the action of a functional in $\left[\mathbb{L}^2(\Omega)\right]^*$ with an arbitrary function (in $\mathbb{L}^2(\Omega)$) is the same as the inner product of its Riesz representation with that arbitrary function. From the RKHS point of view, we have seen from Example 24.1 and Example 24.3, they are different. In particular, the dual space $\left[\mathbb{L}^2(\Omega)\right]^*$ is an RKHS on $\mathbb{L}^2(\Omega)$, but $\mathbb{L}^2(\Omega)$ is not an RKHS on Ω. Of course, if we consider $\mathbb{L}^2(\Omega)$ as the set of linear and continuous functionals on $\left[\mathbb{L}^2(\Omega)\right]^*$, then by a similar reasoning clearly $\mathbb{L}^2(\Omega)$ is a RKHS on $\left[\mathbb{L}^2(\Omega)\right]^*$ (see Problem 24.2).

Example 24.4 ($\mathbb{H}^1[0,1]$ is a RKHS). As we have seen from Example 24.1, that $\mathbb{L}^2[0,1]$ is not a RKHS. It is natural to ask if some of its subspace is a RKHS. A natural candidate is $\mathbb{H}^1[0,1]$ since Definition 13.15 and Lemma 13.3 show that $\mathbb{H}^1[0,1]$ is a Hilbert subspace of $\mathbb{L}^2[0,1]$. For convenience, clarity, and connection with ReLU (to be defined) neuron networks, let us consider the following closed and dense subspace of $\mathbb{H}^1[0,1]$, and thus still Hilbert,

$$\mathscr{H} := \mathbb{H}_0^1[0,1] := \left\{ f \in \mathbb{H}^1[0,1] : f(0) = 0 \right\},$$

with the standard inner product (see (13.14))

$$(f,g)_{\mathscr{H}} := (f,g)_{\mathbb{L}^2} + (\mathscr{D}f, \mathscr{D}g)_{\mathbb{L}^2} := \int_0^1 \overline{f(x)} g(x)\, dx + \int_0^1 \overline{\mathscr{D}f(x)} \mathscr{D}g(x)\, dx$$

where $\mathscr{D}$ is the weak derivative (see Definition 13.14), and the induced norm (see (13.15))

$$\|f\|_{\mathscr{H}} = \sqrt{(f,f)_{\mathscr{H}}}.$$

From Example 13.12, we know that any function f in $\mathscr{H}$: i) has square integrable first-order weak derivative, and ii) is continuous and obeys the fundamental theorem of calculus:

$$f(x) = \int_0^x \mathscr{D}f(t)\, dt. \tag{24.3}$$

Again, since f is equal to its unique continuous representation almost everywhere, we do not distinguish f with the continuous representation. The pointwise evaluation $f(x)$ thus makes sense and we have

$$|\mathsf{e}_x(f)| = |f(x)| \le \|\mathscr{D}f\|_{\mathbb{L}^2} \sqrt{x} \le \sqrt{x}\, \|f\|_{\mathscr{H}}, \tag{24.4}$$

where we have used the Cauchy-Schwarz (13.4) in the first inequality and the definition of the $\mathscr{H}$-norm in the second inequality. Thus the operator norm of e_x is bounded above by $\sqrt{x}$, that is, e_x is thus bounded. By definition, $\mathscr{H}$ is RKHS. The fact that the pointwise evaluation functional e_x is consistent with section 13.2, and we know that e_x is in fact the Dirac delta distribution in this case.

What remains is to determine the two-point kernel function. From the reproducing property (24.1), we have

$$f(x) = (\mathsf{K}_x, f)_{\mathscr{H}} = \left\langle \mathsf{K}_x - \mathscr{D}^2 \mathsf{K}_x, f \right\rangle, \quad \forall f \in \mathscr{H},$$

where $\mathscr{D}$ is again the distributional derivative in Definition 13.12. Since $\mathcal{D}(0,1)$ is dense in $\mathscr{H}$, we have

$$\phi(x) = \left\langle \mathsf{K}_x - \mathscr{D}^2 \mathsf{K}_x, \phi \right\rangle, \quad \forall \phi \in \mathcal{D}(0,1),$$

and by Definition 13.11, we have

$$\mathsf{K}_x(y) - \mathscr{D}^2\mathsf{K}_x(y) = \delta(x-y), \quad \text{and } \mathsf{K}_x(0) = 0,$$

which shows that the two-point kernel is the Green function for the following differential equation with some forcing term $g(y)$

$$u(y) - \mathscr{D}^2 u(y) = g(y), \quad \text{and } u(0) = 0,$$

and this is the direct consequence of the reproducing property (24.1). As can be seen, the two-point kernel $\mathsf{K}(x,y)$—the Green function—depends on the inner product in $\mathscr{H}$. Since we are interested in a special kernel, we now look at a different inner product in $\mathscr{H}$. To that end, we note that a direct consequence (24.4) is the following Friedrichs-Poincaré inequality

$$\sqrt{2}\,\|f\|_{\mathbb{L}^2} \leq \|\mathscr{D}f\|_{\mathbb{L}^2},$$

which implies

$$\sqrt{\frac{2}{3}}\,\|f\|_{\mathscr{H}} \leq \|\mathscr{D}f\|_{\mathbb{L}^2} \leq \|f\|_{\mathscr{H}},$$

that is the $\mathbb{L}^2$-norm of $\mathscr{D}f$ is equivalent to the $\mathscr{H}$-norm. As a result, we can define a new inner product[3] in $\mathscr{H}$ as

$$(f,g)_{\mathscr{H}} := (\mathscr{D}f, \mathscr{D}g)_{\mathbb{L}^2}, \quad \forall f,g \in \mathscr{H}. \tag{24.5}$$

From (24.4), it is also clear that the pointwise evaluation function e_x is continuous and bounded by $\sqrt{x}$. Thus, $\mathscr{H}$ with this new inner product is indeed a RKHS by Definition 24.1. The two-point kernel $\mathsf{K}(x,y)$ is now the Green function of the following Laplace equation with $g(y)$ as some forcing term:

$$-\mathscr{D}^2 u(y) = g(y), \quad \text{and } u(0) = 0.$$

From Example 13.6, we know that $\mathscr{D}\mathsf{K}_x$ is the Heaviside function

$$\mathscr{D}\mathsf{K}_x(y) = \begin{cases} 1 & \text{if } y < x \\ 0 & \text{otherwise} \end{cases}.$$

It follows that

$$\mathsf{K}(y,x) = \int_0^y \mathscr{D}\mathsf{K}_x(z)\,dz = \begin{cases} y & \text{if } y < x \\ x & \text{if } y \geq x \end{cases} = \min\{x,y\},$$

and from (24.2) we have

[3] It is a straightforward exercise to verify the conditions of the inner product in Chapter 4.

$$\|\mathrm{e}_x\|_{\mathscr{H}^*} = \sqrt{\mathsf{K}(x,x)} = \sqrt{x},$$

which is consistent with (24.4). Of course, we can also explicitly verify the reproducing property (24.1) (see Problem 24.4). The connection with ReLU neural network will be discussed in Example 24.5.

We next discuss several important results for an RKHS and its reproducing kernel. We start with an important result in Hilbert space.

Lemma 24.1. *Let $\mathcal{S}$ be a linear subspace of a Hilbert space $\mathscr{H}$. Then, $\mathcal{S}^\perp = \{\theta\}$ iff $\mathcal{S}$ is dense in $\mathscr{H}$.*

Proof. Suppose $\mathcal{S}$ is dense in $\mathscr{H}$ and let $x \in \mathcal{S}^\perp$. We need to show that $\|x\|_{\mathscr{H}}$ is arbitrarily small. Indeed, by the Pythagorean identity (7.3), we have

$$\|x-y\|_{\mathscr{H}}^2 = \|x\|_{\mathscr{H}}^2 + \|y\|_{\mathscr{H}}^2 \geq \|x\|_{\mathscr{H}}^2, \quad \forall y \in \mathcal{S}.$$

Since $\mathcal{S}$ is dense in $\mathscr{H}$, for every $\varepsilon > 0$, we can choose $y \in \mathcal{S}$ such that

$$\|x-y\|_{\mathscr{H}} < \varepsilon \implies \|x\|_{\mathscr{H}} \leq \varepsilon.$$

Since ε is arbitrarily small, we conclude that $\|x\| = 0$, and thus $x = \theta$.

Conversely, suppose $\mathcal{S}^\perp = \{\theta\}$. From Corollary 5.1, we have

$$\overline{\mathcal{S}} = \left(\mathcal{S}^\perp\right)^\perp = \{\theta\}^\perp = \mathscr{H},$$

which, together with Definition 12.1, concludes that $\mathcal{S}$ is a dense subspace of $\mathscr{H}$.

Lemma 24.2 (Density). *Let $\mathscr{H}$ be a RKHS over a set X with values in $\mathbb{F}$ and define*

$$\mathbb{V} := \mathit{span}\{\mathsf{K}_x : x \in X\}.$$

Then $\mathbb{V}$ is dense in $\mathscr{H}$ with respect to the $\mathscr{H}$-norm topology.

Proof. From Lemma 24.1, we only need to show that $\mathbb{V}^\perp = \{\theta\}$. Let $f \in \mathbb{V}^\perp$, we have, by the definition of orthogonality and the reproducing property (24.1),

$$f(x) = (\mathsf{K}_x, f)_{\mathscr{H}} = 0, \quad \forall x \in X,$$

and thus $f = \theta$.

Remark 24.3. While functions in a RKHS space $\mathscr{H}$ could be abstract, functions in the dense subspace $\mathbb{V}$ are simply a linear combination of a finite number of functions of the form K_{x^i} with $i = 1, \ldots, n$ and $n \in \mathbb{N}$. Once the two-point kernel $\mathsf{K}(x,y)$ is determined, $\mathbb{V}$ is explicitly defined. For practical and computational purposes, working with the dense subspace $\mathbb{V}$ is sufficient.

Example 24.5. We continue Example 24.4. The dense subspace in this case reads

$$\mathbb{V} := \operatorname{span}\left\{\mathsf{K}_x\left(y\right) = \min\left\{x, y\right\} : x \in [0,1]\right\}.$$

We now use the density of $\mathbb{V}$ in $\mathscr{H}$ to justify why the set of ReLU neural networks can approximate any functions in $\mathscr{H}$ to any desired accuracy. To that end, let us recall the definition of Rectified Linear Unit (ReLU) [107]:

$$\operatorname{ReLU}\left(x\right) = \min\left\{x, 0\right\}, \quad \forall x \in \mathbb{R}.$$

Now notice that we can write

$$\mathsf{K}_x\left(y\right) = \min\left\{x, y\right\} = \frac{x+y}{2} - \frac{|x-y|}{2} =$$
$$\frac{1}{2}\operatorname{ReLU}\left(x+y\right) - \frac{1}{2}\operatorname{ReLU}\left(-x-y\right) - \frac{1}{2}\operatorname{ReLU}\left(x-y\right) - \frac{1}{2}\operatorname{ReLU}\left(y-x\right),$$

that is, $\mathsf{K}_x\left(y\right)$ is a linear combination of four ReLUs. We also know that single hidden layer ReLU neural networks are linear combinations of a finite number of ReLUs. Thus, $\mathbb{V}$ is a subset of the set of all single hidden layer ReLU networks. Since the former is dense in $\mathscr{H}$, so is the latter.

The next result shows that a closed linear subspace of an RKHS is also an RKHS with a well-defined kernel.

Lemma 24.3. *Let $\mathscr{H}$ be an RKHS over a set X with values in $\mathbb{F}$ and the reproducing kernel K, and $\mathcal{S}$ be a closed linear subspace of $\mathscr{H}$. Then, $\mathcal{S}$ is also an RKHS with the kernel $\tilde{\mathsf{K}}$ such that*

- $\tilde{\mathsf{K}}_x = \mathscr{P}\mathsf{K}_x$
- $\tilde{\mathsf{K}}\left(x,y\right) = \left(\mathscr{P}\mathsf{K}_x, \mathscr{P}\mathsf{K}_y\right)_{\mathscr{H}} = \left(\mathscr{P}\mathsf{K}_x, \mathsf{K}_y\right)_{\mathscr{H}} = \left(\mathsf{K}_x, \mathscr{P}\mathsf{K}_y\right)_{\mathscr{H}}$,

where $\mathscr{P}$ is the orthogonal projection from $\mathscr{H}$ onto $\mathcal{S}$.

Proof. Recall that a closed subspace of a Hilbert space is also Hilbert with the same inner product. Thus, $\mathcal{S}$ is Hilbert. We have

$$\left\|\mathrm{e}_x\right\|_{\mathcal{S}^*} = \sup_{f\in\mathcal{S}} \frac{\left|\mathrm{e}_x\left(f\right)\right|}{\left\|f\right\|_{\mathscr{H}}} \leq \sup_{f\in\mathscr{H}} \frac{\left|\mathrm{e}_x\left(f\right)\right|}{\left\|f\right\|_{\mathscr{H}}} = \left\|\mathrm{e}_x\right\|_{\mathscr{H}^*} < \infty,$$

and thus the pointwise evaluation e_x is also bounded on the subset $\mathcal{S}$. Thus $\mathcal{S}$ is an RKHS. The existence and uniqueness of $\mathscr{P}$ is from Theorem 7.2. By the Riesz representation Theorem 5.1 applying to $\mathcal{S}$ and $\mathscr{H}$ we have

$$\mathrm{e}_x\left(g\right) = \left(g, \tilde{\mathsf{K}}_x\right)_{\mathscr{H}} = \left(g, \mathsf{K}_x\right)_{\mathscr{H}}$$
$$= \left(\mathscr{P}g, \mathsf{K}_x\right)_{\mathscr{H}} = \left(g, \mathscr{P}^*\mathsf{K}_x\right)_{\mathscr{H}} = \left(g, \mathscr{P}\mathsf{K}_x\right)_{\mathscr{H}}, \quad \forall g \in \mathcal{S},$$

which implies $\tilde{\mathsf{K}}_x = \mathscr{P}\mathsf{K}_x$. Thus,

$$\tilde{\mathsf{K}}\left(x,y\right) = \left(\tilde{\mathsf{K}}_x, \tilde{\mathsf{K}}_y\right)_{\mathscr{H}} = \left(\mathscr{P}\mathsf{K}_x, \mathscr{P}\mathsf{K}_y\right)_{\mathscr{H}} = \left(\mathscr{P}\mathsf{K}_x, \mathsf{K}_y\right)_{\mathscr{H}} = \left(\mathsf{K}_x, \mathscr{P}\mathsf{K}_y\right)_{\mathscr{H}},$$

where we have used the fact that $\mathscr{P}^* = \mathscr{P}$ and $\mathscr{P}\mathscr{P} = \mathscr{P}$.

Example 24.6 (Subspace kernel). Recall from Example 24.4 that

$$\mathscr{H} = \mathbb{H}_0^1[0,1] := \left\{ f \in \mathbb{H}^1[0,1] : f(0) = 0 \right\}$$

is an RKHS with the reproducing kernel $\mathsf{K}(x,y) = \min\{x,y\}$. Now consider a closed linear subspace

$$\mathcal{S} := \{ f \in \mathscr{H} : f(1) = 0 \} =: \mathscr{H}_{00}^1[0,1].$$

Clearly, one can follow similar steps as in Example 24.4 to directly show that $\mathscr{H}_{00}^1[0,1]$ is indeed an RKHS and its reproducing kernel is given by

$$\tilde{\mathsf{K}}(x,y) = \begin{cases} (1-y)\,x & \text{if } x \le y \\ (1-x)\,y & \text{if } x \ge y \end{cases}. \tag{24.6}$$

Here we can also follow an indirect approach by exploiting Lemma 24.3. In particular, we immediately know that $\mathscr{H}_{00}^1[0,1]$ is an RKHS with the $\mathscr{H}$ inner product. In order to determine the kernel $\tilde{\mathsf{K}}$ of $\mathscr{H}_{00}^1[0,1]$ we need to find the orthogonal projection onto $\mathscr{H}_{00}^1[0,1]$. From the proof of Corollary 16.1, we simply find an orthogonal basis for $\mathscr{H}_{00}^1[0,1]$, which we essentially already computed in Example 16.4. Indeed, since $v \in \{1\} \cup \left\{ \sqrt{2}\cos(2n\pi x), \sqrt{2}\sin(2n\pi x) : n \in \mathbb{N} \right\}$ is an orthonormal basis vector for $\mathbb{L}^2(0,1)$ and the inner product of $\mathscr{H}$ is defined as in (24.5), orthonormal basis vectors for $\mathscr{H}_{00}^1[0,1]$ can be found by integrating v with zero boundary conditions at 0 and 1. In particular, we have

$$\mathbb{V} := \left\{ \frac{\sqrt{2}}{2n\pi}\sin(2n\pi x), \frac{\sqrt{2}}{2n\pi}(\cos(2n\pi x) - 1) : n \in \mathbb{N} \right\}$$

is an orthonormal basis for $\mathscr{H}_{00}^1[0,1]$ (see Problem 24.5). From Corollary 16.1 we have that[4]

$$\mathscr{P} := \sum_{v_n \in \mathbb{V}} (v_n, \cdot)_{\mathscr{H}}\, v_n$$

is the orthogonal projection onto $\mathscr{H}_{00}^1[0,1]$. From Lemma 24.3, the kernel function for $\mathscr{H}_{00}^1[0,1]$ is then given by

$$\tilde{\mathsf{K}}(x,y) = \tilde{\mathsf{K}}_y(x) = \mathscr{P}\mathsf{K}_y(x) = \sum_{v_n \in \mathbb{V}} (v_n, \mathsf{K}_y)_{\mathscr{H}}\, v_n(x) = \sum_{v_n \in \mathbb{V}} \overline{v_n(y)} v_n(x),$$

which does not give us the explicit form of $\tilde{\mathsf{K}}_y(x)$. However, from the derivation and Definition 24.2, we know that the two-point kernel $\tilde{\mathsf{K}}_y(x)$ is unique

[4] Again, from Corollary 16.1, the convergence of the series is in the strong sense.

(thanks to the Riesz representation Theorem 5.1), and thus $\tilde{\mathsf{K}}_y(x)$ must be the same as the one in (24.6). In other words, the second approach gives us a Fourier series expansion of the reproducing kernel in (24.6). *More importantly, the beauty of this approach is that we immediately yield the Mercer-type theorem for* $\mathscr{H}_{00}^1[0,1]$. From the derivation, we also see that

$$\tilde{\mathsf{K}}(x,y) = \sum_{v_n \in \mathbb{V}} \overline{v_n(y)} v_n(x),$$

holds even without knowing the explicit expression of v_n and without $\mathsf{K}(x,y)$, as long as $\{v_n\}_{n=1}^{\infty}$ is an orthonormal basis for $\mathscr{H}_{00}^1[0,1]$.

Lemma 24.4 (Kernel expansion in terms of orthonormal basis). *Suppose* $\{v_n\}_{n=1}^{\infty}$ *is a countable*[5] *orthonormal basis of a RKHS* $\mathscr{H}$. *Then*

$$\mathsf{K}(x,y) = \sum_{n=1}^{\infty} v_n(x) \overline{v_n(y)},$$

where, for a given y *(and similarly a given* x*), the series converges in the* $\mathscr{H}$*-norm and pointwise.*

Proof. The proof for the expression of K and the $\mathscr{H}$-norm convergence are given in Example 24.6. The pointwise convergence is a direct consequence of Lemma 24.5.

From Example 24.6 we already pointed out that, owing to the Riesz representation Theorem 5.1, each RKHS has a unique reproducing kernel. It turns out that the converse is also true. In particular, if we consider the correspondence between an RKHS and its reproducing kernel as a map, then that map is injective.

Theorem 24.1 (One-to-one correspondence between an RKHS and its reproducing kernel). *Let* $\mathscr{H}^1$ *and* $\mathscr{H}^2$ *are two RKHS on a set* X *with values in* $\mathbb{F}$*, and their corresponding reproducing kernels be* K^1 *and* K^2*. If* $\mathsf{K}^1(x,y) = \mathsf{K}^2(x,y)$ *for all* $x, y \in X$*, then* $\mathscr{H}^1 = \mathscr{H}^2$.

Proof. Since $\mathsf{K}^1(x,y) = \mathsf{K}^2(x,y)$ for all $x, y \in X$, the corresponding dense subsets $\mathbb{V}^1$ and $\mathbb{V}^2$ in Lemma 24.2 are identical and let us denote them as $\mathbb{V}$. Let $f \in \mathbb{V}$, then there exists $n \in \mathbb{N}$, $\{y^i\}_{i=1}^n \subset X$, and $\boldsymbol{\alpha} \in \mathbb{F}^n$ such that

$$f(x) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \mathsf{K}_{y^i}^1(x) = \sum_{i=1}^{n} \boldsymbol{\alpha}_i \mathsf{K}_{y^i}^2(x).$$

Thus

[5] The extension to uncountable cases is possible but it is of no interest in this book.

$$\|f\|_{\mathscr{H}^1}^2 = \left(\sum_{i=1}^n \boldsymbol{\alpha}_i \mathsf{K}_{y^i}^1(x), \sum_{i=1}^n \boldsymbol{\alpha}_i \mathsf{K}_{y^i}^1(x)\right)_{\mathscr{H}^1} = \sum_{i,j=1}^n \overline{\boldsymbol{\alpha}_i} \mathsf{K}^1\left(y^j, y^i\right) \boldsymbol{\alpha}_j$$
$$= \sum_{i,j=1}^n \overline{\boldsymbol{\alpha}_i} \mathsf{K}^2\left(y^j, y^i\right) \boldsymbol{\alpha}_j = \|f\|_{\mathscr{H}^2}^2,$$

which shows that both $\mathscr{H}^1$-norm and $\mathscr{H}^2$-norm are the same and let us denote them as $\|\cdot\|$. In addition, owing to Lemma 24.2, both $\mathscr{H}^1$ and $\mathscr{H}^2$ are the completions of $\mathbb{V}$ under the same norm $\|\cdot\|$, and since the completion is unique (see Remark 5.6), $\mathscr{H}^1$ and $\mathscr{H}^2$ must be identical, and this concludes the proof (see Problem 24.6 for a direct proof that exploits Lemma 24.5).

We next show that in an RKHS $\mathscr{H}$ norm convergence implies pointwise convergence. This is not surprising as the reproducing property (24.1) (which is the direct consequence of the boundedness of the pointwise evaluation functional).

Lemma 24.5 (norm convergence implies pointwise convergence). *Let $\mathscr{H}$ be an RKHS over a set X with values in $\mathbb{F}$, and sequence $\{f_n\}_{n=1}^\infty$. Then,*

$$\|f_n - f\|_{\mathscr{H}} \xrightarrow{n\to\infty} 0 \implies f_n(x) \xrightarrow{n\to\infty} f(x), \forall x \in X.$$

Proof. Owing to the linearity of e_x and the reproducing property (24.1) we have

$$|f_n(x) - f(x)| = |\mathrm{e}_x(f_n - f)| = |(f_n - f, \mathsf{K}_x)_{\mathscr{H}}|$$
$$\leq \|f_n - f\|_{\mathscr{H}} \|\mathsf{K}_x\|_{\mathscr{H}} = \|f_n - f\|_{\mathscr{H}} \sqrt{\mathsf{K}(x,x)} \xrightarrow{n\to\infty} 0,$$

where we have used the Cauchy-Schwarz inequality in $\mathscr{H}$.

Example 24.7. We continue Example 24.5, but now focus on a subspace $\mathcal{C}_0[0,1] \underset{\text{dense}}{\subset} \mathbb{H}_0^1(0,1)$, where the density is a consequence of (13.3). Thus, $\mathbb{V}$, and hence the set of all single hidden layer ReLU neural networks, is also dense in $\mathcal{C}_0[0,1]$ with respect to the $\mathscr{H}$-norm. The result in Lemma 24.5, when taking $f \in \mathcal{C}_0[0,1]$, implies

$$\|f_n - f\|_\infty \leq \|f_n - f\|_{\mathscr{H}} \sqrt{\mathsf{K}(x,x)} \xrightarrow{n\to\infty} 0,$$

where the uniform norm is defined in Definition 13.5. We conclude that $\mathbb{V}$, and hence the set of all single hidden layer ReLU neural networks, is also dense in $\mathcal{C}_0[0,1]$ with respect to the uniform norm.

24.2 From a kernel to its Reproducing Kernel Hilbert Space

We have seen in Theorem 24.1 that every RKHS induces a unique reproducing kernel. We now walk in the opposite direction. Namely, given a kernel, does it induce an RKHS? If it does, is the RKHS unique? But before addressing these questions, we have to answer the more basic question: what is a kernel? It turns out that the two properties **K1)** and **K2)** are sufficient.

Definition 24.3 (Kernel). Let X be a set and $\mathbb{F}$ be a field. A function $\mathsf{K}(\cdot,\cdot) : X \times X \to \mathbb{F}$ is called a kernel if it is symmetric as in **K1)** and symmetric positive definite as in **K2)**. We also define $\mathsf{K}_y(x) := \mathsf{K}(x,y)$.

Theorem 24.2 (One-to-one correspondence between a kernel and its RKHS). *Let K be a kernel as defined in Definition 24.3. Then, three is a unique Hilbert space $\mathscr{H}$ with the properties*

1. $\mathbb{V} := span\{\mathsf{K}_x : x \in X\}$ *is dense in* $\mathscr{H}$, *and*
2. *for any* $x \in X$ *and* $f \in \mathscr{H}$, *the reproducing property* $f(x) = (\mathsf{K}_x, f)_{\mathscr{H}}$ *holds.*

Proof. We shall construct $\mathscr{H}$ and show that it is unique. To that end let us define an inner product in $\mathbb{V}$. For any $f, g \in \mathbb{V}$, there are two indices $n_f, n_g \in \mathbb{N}$, $\{y^i\}_{i=1}^{n_f} \subset X$, and $\{z^j\}_{j=1}^{n_g} \subset X$ such that

$$f(x) = \sum_{i=1}^{n_f} \alpha_i \mathsf{K}_{y^i}(x), \text{ and } g(x) = \sum_{j=1}^{n_g} \beta_i \mathsf{K}_{z^j}(x).$$

An inner product for $\mathbb{V}$ for the two arbitrary f and g can be defined as follows

$$(f,g)_{\mathbb{V}} := \sum_{i=1}^{n_f} \sum_{j=1}^{n_g} \overline{\alpha_i} \mathsf{K}(y^i, z^j) \beta_j,$$

which implies $(\mathsf{K}_x, \mathsf{K}_y)_{\mathbb{V}} = \mathsf{K}(x,y) = \mathsf{K}_y(x)$. We can check that $(f,g)_{\mathbb{V}}$ satisfies all the conditions of inner product in Chapter 4 (see Problem 24.7). Let us denote $\mathscr{H}$ as the completion of $\mathbb{V}$ in the induced norm $\|\cdot\|_{\mathbb{V}}$. We need to verify that $\mathscr{H}$ satisfies the two asserted properties. The first one is obvious by the definition of the closure. For the second property, we first note that from the definition of the inner product in $\mathbb{V}$, the reproducing property holds for all $f \in \mathbb{V}$. Next, for any $f \in \mathscr{H}$, there exists $\{f_n\}_{n=1}^{\infty} \subset \mathbb{V}$ converging to f in the $\mathbb{V}$-norm. The key to notice that is that f can be written in the following form[6]

[6] This can be seen by considering $f_n(x)$ as the nth partial of a series in $\mathbb{V}$. Clearly, such an nth partial sum can be formed as

$$f(x) = \sum_{n=1}^{\infty} \alpha_n \mathsf{K}_{x^n}(x),$$

where the series converges in the $\mathbb{V}$-norm. Owing to the continuity of the inner product, we have

$$(\mathsf{K}_x, f)_{\mathbb{V}} = \sum_{n=1}^{\infty} \alpha_n (\mathsf{K}_x, \mathsf{K}_{x^n})_{\mathbb{V}} = \sum_{n=1}^{\infty} \alpha_n \mathsf{K}_{x^n}(x) = f(x),$$

and thus any $f \in \mathscr{H}$ has the desired reproducing property. To see the uniqueness of $\mathscr{H}$, suppose $\tilde{\mathscr{H}}$ is another Hilbert space satisfying the two asserted properties. Then, both $\mathscr{H}$ and $\tilde{\mathscr{H}}$ are RKHSs by Definition 24.1 and they share the same kernel $\mathsf{K}(x, y)$. By Theorem 24.1, $\mathscr{H} = \tilde{\mathscr{H}}$, and this concludes the proof.

Corollary 24.1. *The correspondence between kernels and their corresponding RKHSs is bijective.*

Proof. This is a direct consequence of Theorem 24.1 and Theorem 24.2.

Example 24.8. Let us now consider all above examples where we have found the kernels explicitly, we know that if we start from these kernels, then the corresponding RKHSs (constructed as in the proof of Theorem 24.2) with the two properties in Theorem 24.2 are exactly the RKHSs that we started in these examples.

We now discuss a few properties of the induced RKHS of a given kernel.

Proposition 24.1. *Let $\mathsf{K}(\cdot,\cdot)$ be a kernel defined on a set X with values in $\mathbb{F}$ and its corresponding RKHS $\mathscr{H}$. Then*

$$\tilde{\mathsf{K}}(x, y) := \overline{\mathsf{K}(x, y)} = \mathsf{K}(y, x)$$

is also a valid kernel and its corresponding RKHS is

$$\overline{\mathscr{H}} := \left\{ \overline{f} : f \in \mathscr{H} \right\}.$$

Proof. The proof is straightforward and is provided in Problem 24.8.

$$f_n(x) = \sum_{i=2}^{n} \underbrace{(f_i(x) - f_{i-1}(x))}_{g_i(x)},$$

where $g_i \in \mathbb{V}$. Thus, the convergence of f_n to f in the $\mathbb{V}$-norm is the same as the convergence of the series $\sum_{n=2}^{\infty} g_n(x)$ to f in the $\mathbb{V}$-norm. In general, f is a countable infinite sum of functions of the form $\mathsf{K}_{x^n}(x)$, and thus resides outside $\mathbb{V}$.

The next result is important for Mercer kernel and Mercer theorem in section 24.3.

Lemma 24.6. *Let $\mathsf{K}(\cdot,\cdot)$ be a kernel defined on a set X with values in $\mathbb{F}$. Furthermore, assume that K is continuous in both of its arguments with respect to the X-topology, and $\mathsf{K}(x,x)$ is uniformly bounded*[7] *for all $x \in X$. Then, $\mathscr{H}$ is continuously embedded in the space of continuous function $\mathcal{C}(X)$, that is, $\mathscr{H} \xhookrightarrow[continuous]{} \mathcal{C}(X)$.*

Proof. Clearly, all functions in $\mathbb{V}$ defined in Lemma 24.2 are continuous by the assumption. Now let $f \in \mathscr{H}$, then there exists a sequence $\{f_n\}_{n=1}^{\infty} \in \mathbb{V}$ such that $\|f_n - f\|_{\mathscr{H}} \xrightarrow{n\to\infty} 0$. Since the uniform limit of a sequence of continuous functions is continuous, the proof is concluded if we can show $\|f_n - f\|_{\infty} \xrightarrow{n\to\infty} 0$. But this is readily available from the reproducing property (24.1) and the proof is similar to that of Lemma 24.5. Indeed, we have

$$\begin{aligned} |f_n(x) - f(x)| = (\mathsf{K}_x, f_n - f)_{\mathscr{H}} &\le \|\mathsf{K}_x\|_{\mathscr{H}} \|f_n - f\|_{\mathscr{H}} \\ &= \sqrt{\mathsf{K}(x,x)}\, \|f_n - f\|_{\mathscr{H}}. \end{aligned}$$

It follows that

$$\|f_n - f\|_{\infty} = \sup_{x\in X} \sqrt{\mathsf{K}(x,x)}\, \|f_n - f\|_{\mathscr{H}} \xrightarrow{n\to\infty} 0.$$

The continuity of the injection $\iota : \mathscr{H} \to \mathcal{C}(X)$ is again due to the reproducing property as

$$\|f\|_{\infty} \le \sup_{x\in X} \sqrt{\mathsf{K}(x,x)}\, \|f\|_{\mathscr{H}},$$

and thus

$$\|\iota\| \le \sup_{x\in X} \sqrt{\mathsf{K}(x,x)} < \infty.$$

24.3 Kernel-based integral operators and the Mercer theorem

Definition 24.4 (Mercer Kernel). If $\mathsf{K} : X \times X \to \mathbb{F}$ is continuous and satisfies Definition 24.3 is called a Mercel kernel.

For this section, we assume that the kernel $\mathsf{K} : X \times X \to \mathbb{F}$ is *Mercer*. As a consequence of Lemma 24.6, the first interesting property of a Mercer kernel is that its unique induced RKHS is a subspace of the space of continuous functions on X. In this section, we further explore other properties RKHS induced from Mercer kernels.

[7] A sufficient condition is that X is compact, as then by the Weierstrass theorem [12, 87, 142] we have that $\sup_{x\in X} K(x,x) < \infty$.

24.3.1 "Weak" compactness of a closed and bounded set in an RKHS

We begin by looking at the result of Lemma 24.6 from a topology point of view. Since the injection (the identity map) is continuous, that is, the $\mathscr{H}$ topology must be stronger (have more open sets) than the $\mathcal{C}$ topology. As we shall show, $\mathcal{C}$ topology has so many less open balls that a closed and bounded $\mathscr{H}$ ball is covered by a finite number of $\mathcal{C}$ balls, and thus it is compact in the $\mathcal{C}$ topology. We shall discuss the theoretical and practical implications of this important fact. Let us first recall a well-known result in Hilbert space.

Lemma 24.7 (Weak compactness of closed balls in Hilbert space). *If $\boldsymbol{B}$ is a closed ball in a Hilbert space $\mathscr{H}$, it is weakly compact. In other words, every sequence $\{f_n\}_{n\in\mathbb{N}} \subset \boldsymbol{B}$ has a weakly convergence subsequence $\{f_{n_k}\}_{k\in\mathbb{N}}$. That is, there exists some $f \in \boldsymbol{B}$ such that*

$$\lim_{k\to\infty} (f_{n_k}, g)_{\mathscr{H}} = (f, g)_{\mathscr{H}}, \quad \forall g \in \mathscr{H}.$$

We next show that a closed and bounded $\mathscr{H}$ ball is closed in the $\mathcal{C}$ topology.

Proposition 24.2 (A closed and bounded $\mathscr{H}$ ball is closed in the $\mathcal{C}$ topology). *Suppose K be a Mercer kernel on a compact metric space X with values in $\mathbb{F}$, and $\mathscr{H}$ is the associated RKHS. For any $r > 0$, the ball $\boldsymbol{B}(r) := \{f \in \mathscr{H} : \|f\|_{\mathscr{H}} \leq r\}$ is a closed subset of $\mathcal{C}(X)$.*

Proof. From Theorem Lemma 24.6, it is sufficient to show that $\boldsymbol{B}(r)$ is closed in $\mathcal{C}(X)$. To that end, suppose $\{f_n\}_{n\in\mathbb{N}} \subset \boldsymbol{B}(r)$ converges (in the uniform norm) to $f \in \mathcal{C}(X)$, i.e.,

$$\lim_n f_n(x) = f(x), \forall x \in X.$$

We need to show that $f \in \boldsymbol{B}(r)$. Since, by Lemma 24.7, $\boldsymbol{B}(r)$ is weakly compact, there exists a subsequence $\{f_{n_k}\}_{k\in\mathbb{N}}$ converging to $\hat{f} \in \boldsymbol{B}(r)$, i.e.,

$$\lim_k (f_{n_k}, g)_{\mathscr{H}} = \left(\hat{f}, g\right)_{\mathscr{H}}, \quad \forall g \in \mathscr{H}.$$

Now taking $g = \mathsf{K}_x$ and using the reproducing property (24.1) we have

$$f(x) = \lim_k f_{n_k}(x) = \lim_k (f_{n_k}, \mathsf{K}_x)_{\mathscr{H}} = \left(\hat{f}, \mathsf{K}_x\right)_{\mathscr{H}} = \hat{f}(x), \quad \forall x \in X.$$

Since both f and $\hat{f}$ are continuous, they must be identical and this ends the proof. □

In order to show that $\boldsymbol{B}(r) := \{f \in \mathscr{H} : \|f\|_{\mathscr{H}} \leq r\}$ is compact subset of $\mathcal{C}(X)$, we need to recall the Arzelá-Ascoli theorem.

Definition 24.5 (Equicontinuity). A subset $\mathcal{M}$ of $\mathcal{C}(X)$ is equicontinuous at $x \in X$ if for any $\varepsilon > 0$ there exists a neighborhood $\boldsymbol{B}$ of x such that $\forall t \in X$ and $\forall f \in \mathcal{M}$ we have $\|f(x) - f(t)\|_{\infty} < \varepsilon$. $\mathcal{M}$ is equicontinuous if it is equicontinuous at every $x \in X$.

Theorem 24.3 (Arzelá-Ascoli theorem). *Let X be a compact metric space. $\mathcal{M} \subset \mathcal{C}(X)$ is compact if and only if $\mathcal{M}$ is closed, bounded, and equicontinuous.*

We are in the position to prove the compactness of $\boldsymbol{B}(r)$ in the $\mathcal{C}$ topology.

Theorem 24.4. *Suppose K be a Mercer kernel on a compact metric space X, and $\mathscr{H}$ is the associated RKHS. The inclusion $\iota : \mathscr{H} \hookrightarrow \mathcal{C}(X)$ is compact. In other words, the set $\iota(\boldsymbol{B}(r))$ is compact in $\mathcal{C}(X)$ for any $r > 0$.*

Proof. Lemma 24.6 and Proposition 24.2 show that $\iota(\boldsymbol{B}(r))$ is closed and bounded. From the Arzelá-Ascoli Theorem 24.3, the assertion of the theorem holds if we can show the equicontinuity. We have

$$\begin{aligned} |f(x) - f(t)| &= |(f, \mathsf{K}_x - \mathsf{K}_t)_{\mathscr{H}}| \leq \|f\|_{\mathscr{H}} \|\mathsf{K}_x - \mathsf{K}_t\|_{\mathscr{H}} \\ &= r\sqrt{(\mathsf{K}_x - \mathsf{K}_t, \mathsf{K}_x - \mathsf{K}_t)_{\mathscr{H}}} = r\sqrt{\mathsf{K}_x(x) - \mathsf{K}_x(t) + \mathsf{K}_t(t) - \mathsf{K}_t(x)} \end{aligned}$$

Now since K is continuous on the compact set $X \times X$, it is uniformly continuous on $X \times X$, i.e., for any $\varepsilon > 0$, there exists $\delta > 0$ such that for all $x, t, t' \in X$, $\|t - t'\|_X < \delta$ (δ does not depend on x, t, t') implies

$$|\mathsf{K}_x(t) - \mathsf{K}_x(t')| < \varepsilon.$$

We thus have

$$|f(x) - f(t)| \leq r\sqrt{2\varepsilon}, \quad \forall t, x : \|x - t\| \leq \delta, \quad \forall f \in \iota(\boldsymbol{B}(r)),$$

and this concludes the proof.

Theoretically, this result is interesting as in infinite dimensional Hilbert space, a closed and bounded ball is typically not compact (in the norm topology). Recall Banach-Alaoglu theorem [112, 109, 28, 12, 87, 142] applied to Hilbert spaces that any closed and bounded ball (in the norm topology) is weakly compact. In the same spirit, Theorem 24.4 says that we can view $\mathcal{C}$ topology as a weak topology, in the RKHS generated by a Mercer kernel, in which any bounded sequence in the RKHS contains a convergent subsequence. Such a sequential weakly compactness is important to the existence of optimization[8] on a closed and bounded set of an RKHS. In fact, it is what we shall explore in the representer Theorem 24.5.

[8] Instead of the requirement on continuity of a function on a compact set for the existence of a minimizer, a much weaker condition can be deployed: A weakly lower semi-continuous

24.3.2 A representer theorem

Theorem 24.5 (Representer theorem). *Let $\mathscr{H}$ be a RKHS generated by Mercer kernel K on a set X and let $J : \mathbb{R}^n \to \mathbb{R}$ be any continuous function. Then the following minimization problems are equivalent.*

i)

$$\min_{f \in \mathscr{H},\, \|f\|_{\mathscr{H}} \leq \lambda} J\left(f\left(x^1\right), ..., f\left(x^n\right)\right),$$

where $0 < \lambda < \infty$.

ii)

$$\min_{f \in \mathbb{V}_n,\, \|f\|_{\mathscr{H}} \leq \lambda} J\left(f\left(x^1\right), ..., f\left(x^n\right)\right)$$

where $\mathbb{V}_n = span\left\{ f \in \mathscr{H} : f(x) = g_{\boldsymbol{\alpha}}(x) = \sum_{i=1}^{n} \alpha_i \mathsf{K}_{x^i}(x) \right\}$

iii)

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n,\, \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha} \leq \lambda^2} J\left(g_{\boldsymbol{\alpha}}\left(x^1\right), ..., g_{\boldsymbol{\alpha}}\left(x^n\right)\right)$$

where $\boldsymbol{K}_{ij} = \mathsf{K}\left(x^i, x^j\right)$.

Note that J, as a function of f, is continuous since the pointwise evaluation functional is continuous in f by Definition 24.1. Thanks to Theorem 24.4 the first optimization is thus a minimization of a continuous function over a compact subset of $\mathcal{C}(X)$ and thus has at least one minimizer by Weierstrass theorem [112, 109, 28, 12, 87, 142].

Proof. In this proof, all inner products will be assumed to be the inner product on the RKHS and all norms are the norms induced by this inner product unless otherwise specified. That is, $\|\cdot\| = \|\cdot\|_{\mathscr{H}}$.

For any $f \in \mathscr{H}$, from Theorem 7.2 and Lemma 7.2, there exists a unique decomposition

$$f = \bar{f} + f^{\perp}$$

where $\bar{f} \in \mathbb{V}_n$ is the orthogonal projection of f in $\mathbb{V}_n$ and $f^{\perp} \in \mathbb{V}_n^{\perp}$. By the Pythagorean (7.3), we have

$$\|f\|^2 = \left\|\bar{f}\right\|^2 + \left\|f^{\perp}\right\|^2.$$

Now since $f^{\perp} \in \mathscr{H}$, the reproducing property gives

$$f^{\perp}\left(x^i\right) = \left(\mathsf{K}_{x^i}, f^{\perp}\right) = 0,$$

where the second equality is from the fact that $\mathsf{K}_{x^i} \in \mathbb{V}_n$ and $f^{\perp} \in \mathbb{V}_n^{\perp}$. We conclude that

on a weakly compact set possesses an infimizer [87, 142]. However, we do not dwell on this subject as it is not within the scope of this adjoint book.

$$f\left(x^i\right) = \bar{f}\left(x^i\right).$$

Therefore, we can remove the dependence on $f^\perp$ from the constraint, proving the equivalency of the first and second optimization problems.

For the second equivalency (ii $\Leftrightarrow$ iii), it is trivial to replace the optimization constraint of $f \in \mathbb{V}_n$ with $\boldsymbol{\alpha} \in \mathbb{R}^n$ using the definition of $\mathbb{V}_n$. What remains is to show

$$\|g_{\boldsymbol{\alpha}}\|^2 = \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha},$$

but this is obvious as

$$\begin{aligned}\|g_{\boldsymbol{\alpha}}\|^2 &= (g_{\boldsymbol{\alpha}}, g_{\boldsymbol{\alpha}}) \\ &= \sum_{i,j=1}^{n} \alpha_i \alpha_j \mathsf{K}\left(x^i, x^j\right) \\ &= \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha}.\end{aligned}$$

Lemma 24.8 (Minimum norm interpolant). *Let $\mathscr{H}$ be a RKHS generated by positive definite Mercer kernel K on a set X. Given a set of data points $\left\{x^j, y^j\right\}_{j=1}^n$. Then, the following optimization problem has a unique minimizer f^* given as*

$$\begin{aligned}&\min_{f \in \mathscr{H}} \|f\|_{\mathscr{H}}^2 \\ &\text{subject to } f\left(x^j\right) = y^j, \quad j = 1, \ldots, n\end{aligned}$$

$$f^*(x) = \sum_{j=1}^{n} \alpha_j \mathsf{K}_{x^j}(x),$$

where $\boldsymbol{\alpha} = \boldsymbol{K}^{-1}\boldsymbol{y}$, with $\boldsymbol{K}_{ij} = \mathsf{K}\left(x^i, x^j\right)$, $\boldsymbol{y} = \left[y^1, \ldots, y^n\right]^\top$, and $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha^n]^\top$.

Proof. The proof is a variant of the proof of Theorem 24.5. In particular, the optimization can be written equivalently as

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} \boldsymbol{\alpha}^\top \boldsymbol{K} \boldsymbol{\alpha}^\top + \left\|f^\perp\right\|_{\mathscr{H}}^2 \text{ subject to } \boldsymbol{K}\boldsymbol{\alpha} = \boldsymbol{y},$$

where $f = g_{\boldsymbol{\alpha}} + f^\perp$. Clearly, f that has minimum norm when $f^\perp = \theta$, and this ends the proof as there is only one $\boldsymbol{\alpha} = \boldsymbol{K}^{-1}\boldsymbol{y}$, and thus the unique minimizer is $f^*(x) = \sum_{j=1}^{n} \alpha_j \mathsf{K}_{x^j}(x)$.

24.3.3 Kernel-based integral operators and the Mercer theorem

Let us define the following integral operator

$$(\mathcal{L}_{\mathsf{K}} f)(x) := \int_X \mathsf{K}_x(t) f(t)\, d\pi(t) = \int_X \mathsf{K}(t,x) f(t)\, d\pi(t), \tag{24.7}$$

where we assume π is a probability measure on X though all the results, up to a constant, also hold for a general Borel measure. For simplicity, we assume that X is compact and $\mathsf{K} : X \times X \to \mathbb{R}$ is continuous and non-trivial.[9] Thus, K is uniformly continuous as X is compact and this immediately implies that $\mathcal{L}_{\mathsf{K}} f \in \mathcal{C}(X)$ (see Problem 24.9). Let us define (owing to the Weierstrass theorem [112, 109, 28, 12, 87, 142])

$$c_{\mathsf{K}} := \sup_{x,t\in X} |\mathsf{K}(x,t)| < \infty, \tag{24.8}$$

which immediately implies $\|\mathsf{K}_x\|_{\mathbb{L}^2(X,\pi)} \le c_{\mathsf{K}}$, that is, $\mathsf{K}_x \in \mathbb{L}^2(X,\pi)$ for all $x \in X$. By Cauchy-Schwarz inequality (13.4), we have

$$|(\mathcal{L}_{\mathsf{K}} f)(x)| \le \|\mathsf{K}_x\|_{\mathbb{L}^2(X,\pi)} \|f\|_{\mathbb{L}^2(X,\pi)} \le c_{\mathsf{K}} \|f\|_{\mathbb{L}^2(X,\pi)}, \tag{24.9}$$

which implies $\mathcal{L}_{\mathsf{K}}$ as a map from $\mathbb{L}^2(X,\pi)$ to $\mathcal{C}(X)$ is (Lipschitz) continuous. Since the inclusion $\mathcal{C}(X) \hookrightarrow \mathbb{L}^2(X,\pi)$ is continuous, $\mathcal{L}_{\mathsf{K}}$, as a linear map from $\mathbb{L}^2(X,\pi)$ into $\mathbb{L}^2(X,\pi)$, is continuous and its operator norm is bounded as $\|\mathcal{L}_{\mathsf{K}}\| \le c_{\mathsf{K}}$.

Proposition 24.3. *The operator $\mathcal{L}_{\mathsf{K}} : \mathbb{L}^2(X,\pi) \to \mathbb{L}^2(X,\pi)$ defined in (24.7) with condition (24.8) is a compact operator and it is self-adjoint. If, in addition, X is compact, then $\mathcal{L}_{\mathsf{K}}$ is positive semidefinite.*

Proof. Consider the ball $\boldsymbol{B}(r) := \left\{ f \in \mathbb{L}^2(X,\pi) : \|f\|_{\mathbb{L}^2(X,\pi)} \le r \right\}$ in $\mathbb{L}^2(X,\pi)$. We are going to show that the image $\mathcal{L}_{\mathsf{K}}(\boldsymbol{B})$, which is a subset of $\mathcal{C}(X)$ by (24.9), is relatively compact in $\mathcal{C}(X)$. From (24.9) we see that $\mathcal{L}_{\mathsf{K}}(\boldsymbol{B}(r))$ is uniformly bounded. A similar argument as in (24.9) shows that

$$|(\mathcal{L}_{\mathsf{K}} f)(x) - (\mathcal{L}_{\mathsf{K}} f)(x')| \le 2c_{\mathsf{K}} \|f\|_{\mathbb{L}^2(X,\pi)} \le 2c_{\mathsf{K}} r, \quad \forall f \in \boldsymbol{B}(r),$$

which implies that $\mathcal{L}_{\mathsf{K}}(\boldsymbol{B}(r))$ is equicontinuous. By the Arzelá-Ascoli theorem Theorem 24.3, the closure of $\mathcal{L}_{\mathsf{K}}(\boldsymbol{B})$ is compact in $\mathcal{C}(X)$. In other words, $\mathcal{L}_{\mathsf{K}}(\boldsymbol{B}(r))$ is relatively compact and by definition Definition 14.1, $\mathcal{L}_{\mathsf{K}} : \mathbb{L}^2(X,\pi) \to \mathcal{C}(X)$ is a compact operator. Since the inclusion $\mathcal{C}(X) \hookrightarrow$

[9] Considering complex-valued kernel is a straightforward modification and we leave it for the interested readers.

$\mathbb{L}^2(X,\pi)$ is continuous, we conclude that $\mathcal{L}_{\mathsf{K}} : \mathbb{L}^2(X,\pi) \to \mathbb{L}^2(X,\pi)$ is a compact operator.

The self-adjointness is clear by the Fubini theorem (thanks to (24.8)). The positive semi-definiteness of $\mathcal{L}_{\mathsf{K}}$ is a direct consequence of the positive semidefiniteness of K. Indeed, since X is compact, there exists $n \in \mathbb{N}$ such that we can subdivide X into n subsets with equal volumes and with "centroids" $x^1, \ldots, x^n$. We have

$$\begin{aligned}(f, \mathcal{L}_{\mathsf{K}} f)_{\mathbb{L}^2(X,\pi)} &= \int_{X\times X} \mathsf{K}(t,x)\, f(t)\, f(x)\, d\pi(t)\, d\pi(x) \\ &= \lim_{n\to\infty} \frac{\mathcal{V}(X)^2}{n^2} \sum_{i,j=1}^{n} \mathsf{K}(t_j, x_i)\, f(t_j)\, f(x_i) \geq 0, \end{aligned} \tag{24.10}$$

where we have used the positive semi-definiteness of K.

By the Hilbert-Schmidt Theorem 14.1, any $f \in \mathbb{L}^2(X,\pi)$ can be represented as

$$f(x) = \sum_{i=1}^{\infty} (\varphi_i, f)\, \varphi_i(x) + (\mathscr{P} f)(x), \tag{24.11}$$

where $\mathscr{P}$ is the projection of $\mathbb{L}^2(X,\pi)$ onto the nullspace $\mathsf{N}(\mathcal{L}_{\mathsf{K}})$, and the action of $\mathcal{L}_{\mathsf{K}}$ on f can be written as (by linearity and continuity of $\mathcal{L}_{\mathsf{K}}$)

$$\mathcal{L}_{\mathsf{K}} f = (\mathsf{K}_x, f)_{\mathbb{L}^2(X,\pi)} = \sum_{i=1}^{\infty} a_i \lambda_i \varphi_i,$$

with eigenpairs $\{\lambda_i, \varphi_i\}_{i=1}^{\infty}$ such that $\lambda_1 \geq \lambda_2 \geq \ldots > 0$ and $\mathcal{L}_{\mathsf{K}}\varphi_i = (\varphi_i, \mathsf{K}_x)_{\mathbb{L}^2(X,\pi)} = \lambda_i \varphi_i(x)$, and the convergence is in $\mathbb{L}^2(X,\pi)$. Thus, from section 5.4, we see that the (possibly) "infinite diagonal matrix", with $\{\lambda_i\}_{i=1}^{\infty}$ as the diagonal elements, is the representation of K in the orthonormal set $\{\varphi_i\}_{i=1}^{\infty}$. In other words, we can express K as

$$\mathsf{K}(x,y) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x)\, \varphi_i(y), \tag{24.12}$$

where the convergence of the series on the right hand side is in the operator norm (see [109, Theorem 6.11.1]) as K, identified as $\mathcal{L}_{\mathsf{K}}$, is a compact self-adjoint linear operator from $\mathbb{L}^2(X,\pi)$ into $\mathbb{L}^2(X,\pi)$.

Another way to derive the expansion of the kernel in (24.12) is to use the Fourier expansion (24.11). To that end, we note that

$$(\mathcal{L}_{\mathsf{K}} \mathsf{K}_x)(x) = \int_X |\mathsf{K}_x(t)|^2\, d\pi(t) > 0,$$

Since K is non-trivial. This shows $\mathsf{K}_x \notin \mathsf{N}(\mathcal{L}_\mathsf{K})$. Now applying (24.11) for K_x we have

$$\mathsf{K}(x,y) = \mathsf{K}_y(x) = \sum_{i=1}^{\infty} (\varphi_i, \mathsf{K}_y)\,\varphi_i(x) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x)\,\varphi_i(y),$$

where we have used the fact that $\varphi_i(y)$ are eigenfunctions of K. By the Parseval identity (14.3) we have

$$\|\mathsf{K}_y(x)\|^2_{\mathbb{L}^2(X,\pi)} = \sum_{i=1}^{\infty} |(\varphi_i, \mathsf{K}_y)|^2 = \sum_{i=1}^{\infty} |\lambda_i|^2 |\varphi_i(y)|^2,$$

from which it follows that

$$\infty > c_\mathsf{K}^2 > \int_{X\times X} |\mathsf{K}(x,y)|^2 \, d\pi(x)\, d\pi(y) = \sum_{i=1}^{\infty} |\lambda_i|^2 \|\varphi_i\|^2_{\mathbb{L}^2(X,\pi)} = \sum_{i=1}^{\infty} {\lambda_i}^2,$$

where we have interchanged the integral and infinite series in the first equality (thanks to the monotone convergence theorem [112, 109, 28, 12, 87, 142]), and the fact that each eigenvalue λ_i is non-negative (thanks to the positive semidefiniteness of the kernel in (24.10)) in the last equality. That is, the sequence of eigenvalues of K is square summable.

Theorem 24.6 (Mercer theorem). *Let X be compact and consider $\mathcal{L}_\mathsf{K} : \mathbb{L}^2(X,\pi) \to \mathbb{L}^2(X,\pi)$ defined in (24.7) where K satisfies condition (24.8) and is continuous on $X \times X$. Then*

$$\mathsf{K}(x,y) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(x)\,\varphi_i(y),$$

converges in $\mathbb{L}^2(X,\pi)$

Problems

Problem 24.1. This is another example showing that $\mathbb{L}^2(\Omega)$ with $\Omega \subseteq \mathbb{R}^n$ is not a RKHS on Ω. Let $x \in (0,1)$, and consider the following function

$$f_n(y) = \begin{cases} \left(\dfrac{y}{x}\right)^n & \text{if } 0 \le y \le x \\ \left(\dfrac{1-y}{1-x}\right)^{2n} & \text{if } x < y \le 1 \end{cases}$$

Show that $f_n \in \mathbb{L}^2(0,1)$ and that $\mathrm{e}_x(f_n) = 1$ for $x \in (0,1)$. Then show that e_x is an unbounded linear operator on $\mathbb{L}^2(0,1)$.

Hint Using the same strategy as in Example 24.1.

Problem 24.2. Consider $\mathbb{L}^2(\Omega)$ as the set of linear and continuous functionals on $\left[\mathbb{L}^2(\Omega)\right]^*$. Show that $\mathbb{L}^2(\Omega)$ is a RKHS on $\left[\mathbb{L}^2(\Omega)\right]^*$ and determine the kernel function.

Hint. Follow the same steps as Example 24.3

Problem 24.3. The following problems are similar to Example 24.4 but with different boundary conditions. Show that the following spaces are RKSH and find their two-point kernel functions

- Homogeneous Dirichlet boundary conditions

$$\mathscr{H} := \mathbb{H}_0^1[0,1] := \left\{ f \in \mathbb{H}^1[0,1] : f(0) = 0 \text{ and } f(1) = 0 \right\}.$$

- Periodic boundary conditions

$$\mathscr{H} := \mathbb{H}_0^1[0,1] := \left\{ f \in \mathbb{H}^1[0,1] : f(0) = f(1) \right\}.$$

Problem 24.4. Verify the reproducing property for the two-point kernel $\mathsf{K}(x,y) = \min\{x,y\}$ in Example 24.4.

Problem 24.5. Show that

$$\mathbb{V} := \left\{ \frac{\sqrt{2}}{2n\pi} \sin(2n\pi x), \frac{\sqrt{2}}{2n\pi} (\cos(2n\pi x) - 1) : n \in \mathbb{N} \right\},$$

is an orthonormal basis for $\mathcal{S}$ in Example 24.6.

Hint. The orthonormality of $\mathbb{V}$ in the $\mathscr{H}$-inner product is clear. Suppose $f \in \mathcal{S}$ and $(f,u) = 0$ for all $u \in \mathbb{V}$. Then using the $\mathscr{H}$-inner product to conclude that f' must be a constant since it is orthogonal to all non-constant Fourier modes. Thus, f is a linear polynomial, but since $f(1) = f(0) = 0$, we can conclude that $f = \theta$.

Problem 24.6. In Theorem 24.1, we show that $\mathscr{H}^1 = \mathscr{H}^2$ using the uniqueness of the closure. We now achieve the same goal using a direct approach. Let $f \in \mathscr{H}^1$, show that $f \in \mathscr{H}^2$, and vice versa. Thus, $\mathscr{H}^1 = \mathscr{H}^2$.

Hint. Since $\mathbb{V}$ is dense in $\mathscr{H}^1$, there exists $\{f_n\}_{n=1}^{\infty} \subset \mathbb{V}$ and $\lim_{n\to\infty} \|f_n - f\| = 0$. In particular, $\{f_n\}_{n=1}^{\infty}$ is Cauchy in $\mathscr{H}^2$ and let g be the limit in $\mathscr{H}^2$. Now invoking Lemma 24.5 we have

$$g(x) = \lim_{n\to\infty} f_n(x) = f(x),$$

and thus $f \in \mathscr{H}^2$.

Problem 24.7. Show that the inner product defined for $\mathbb{V}$ in the proof of Theorem 24.2 is indeed a valid inner product.

Hint the linearity with respect to the second argument g and the symmetry (a direct consequence of the symmetry of the kernel condition **K1)**) are clear. By the symmetric positive definite condition **K2)** of the kernel, it is also clear that $(f,f)_{\mathbb{V}} \geq 0$. Finally, from $(f,f)=0$ we need to show that $f=\theta$. But this is clear from the Cauchy-Schwarz inequality

$$|f(x)| = |(\mathsf{K}_x, f)_{\mathbb{V}}| \leq \sqrt{(\mathsf{K}_x, \mathsf{K}_x)_{\mathbb{V}}}\sqrt{(f,f)_{\mathbb{V}}}.$$

Problem 24.8. Prove Proposition 24.1.

Problem 24.9. Consider the kernel integral operator defined in (24.7) and suppose that X is compact and $\mathsf{K} : X \times X \to \mathbb{F}$ is continuous. Show that $\mathcal{L}_{\mathsf{K}} f \in \mathcal{C}(X)$.

Hint. The uniform continuity of K allows us to switch the integral and limit.

Chapter 25
The role of adjoint in ADMM

Abstract

Glossary

Use the template *glossary.tex* together with the Springer document class SVMono (monograph-type books) or SVMult (edited books) to style your glossary in the Springer layout.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

glossary term Write here the description of the glossary term. Write here the description of the glossary term. Write here the description of the glossary term.

Solutions

Problems of Chapter ??

?? The solution is revealed here.

?? Problem Heading
(a) The solution of first part is revealed here.
(b) The solution of second part is revealed here.

References

1. Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dan Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.
2. J. Abadie and J. Carpentier. Generalization of the Wolfe reduced gradient method t o the case of nonlinear constraints. Optimization, Sympos. Inst. Math. Appl. Univ. Keele, England, 1968, 37-47 (1969)., 1969.
3. R. A. Adams and J. J. F. Fournier. *Sobolev Spaces*. Elsevier Science, Amsterdam, Boston, Heidelberg,..., second edition, 2003.
4. Mark Ainsworth and J. Tinsley Oden. A posteriori error estimators for the stokes and oseen equations. *SIAM Journal on Numerical Analysis*, 34(1):228–245, 1997.
5. Mark Ainsworth and J. Tinsley Oden. *A Posteriori Error Estimation in Finite Element Analysis*. Wiley, August 2000.
6. Albert Orwa Akuno, L. Leticia Ramirez-Ramirez, Chahak Mehta, Tan Bui-Thanh, and Jose Montoya. Multi-patch epidemic models with partial mobility, residency, and demography. *Chaos, Solitons and Fractals: the interdisciplinary journal of Nonlinear Science, and Nonequilibrium and Complex Phenomena, Accepted*, 2023.
7. C.D. Aliprantis. *An Invitation to Operator Theory*. Graduate studies in mathematics. American Mathematical Society, 2002.
8. Nenad Antonić and Krevsimir Burazin. Graph spaces of first-order linear partial differential operators. *Mathematical Communications*, 14(1):135–155, 2009.
9. Athanasios C. Antoulas. *Approximation of Large-Scale Dynamical Systems*. Society for Industrial and Applied Mathematics, 2005.
10. Tom M Apostol. *Calculus*. John Wiley & Sons, Nashville, TN, 2 edition, June 1967.
11. Todd Arbogast and Jerry L. Bona. *Methods of Applied Mathematics*. University of Texas at Austin, 2008. Lecture notes in applied mathematics.
12. Todd Arbogast and Jerry L Bona. *Functional analysis for the applied mathematician*. Chapman and Hall/CRC, Boca Raton, December 2024.
13. Sheldon Axler. *Linear Algebra Done Right*. Undergraduate Texts in Mathematics. Springer International Publishing, Cham, 3rd ed. edition, 2015.
14. Stefan Banach. *Theorie des Operations Lineaires*. AMS Chelsea Publishing. Chelsea Publishing, Providence, RI, 2 edition, January 1978.
15. Roland Becker and Rolf Rannacher. An optimal control approach to a posteriori error estimation in finite element methods. *Acta Numerica*, 10:1–102, 2001.
16. G Berkooz, P Holmes, and J L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1):539–575, 1993.
17. G Berkooz, P Holmes, and J L Lumley. The proper orthogonal decomposition in the analysis of turbulent flows. *Annual Review of Fluid Mechanics*, 25(1):539–575, 1993.
18. D. P. Bertsekas. *Constrained Optimization and Lagrange Multiplier Methods*. Academic Press, New York,. London, Paris, San Diego, San Francisco, 1982.
19. D. P. Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, 1999.
20. A. Beskos, F. J. Pinski, J. M. Sanz-Serna, and A. M. Stuart. Hybrid Monte Carlo on Hilbert spaces. *Stochastic Processes and their Applications*, 121:2201–2230, 2011.
21. L.T. Biegler, O. Ghattas, M. Heinkenschloss, D. Keyes, and B. Waanders. *Real-time PDE-constrained Optimization*. Computational Science and Engineering. Society for Industrial and Applied Mathematics (SIAM, 3600 Market Street, Floor 6, Philadelphia, PA 19104), 2007.

22. L.T. Biegler, O. Ghattas, M. Heinkenschloss, and B. Waanders. *Large-Scale PDE-Constrained Optimization*. Lecture Notes in Computational Science and Engineering. Springer Berlin Heidelberg, 2012.
23. A. Borzi and V. Schulz. *Computational Optimization of Systems Governed by Partial Differential Equations*. Computational Science and Engineering. Society for Industrial and Applied Mathematics, 2012.
24. James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018.
25. Fred Brauer, Carlos Castillo-Chavez, and Zhilan Feng. *Mathematical Models in Epidemiology*. Springer New York, 2019.
26. Fred Brauer, Pauline van den Driessche, and Jianhong Wu, editors. *Mathematical Epidemiology*. Springer Berlin Heidelberg, 2008.
27. A. Bressan. *Lecture Notes on Functional Analysis: With Applications to Linear Partial Differential Equations*. Graduate studies in mathematics. American Mathematical Society, 2013.
28. H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Universitext. Springer New York, 2010.
29. A. E. Bryson. *Applied Optimal Control: Optimization, Estimation and Control (1st ed.)*. Routledge, 1975.
30. Tan Bui-Thanh. Discretization-invariant active subspace methods. *In preparation*, 2024.
31. Tan Bui-Thanh, Leszek Demkowicz, and Omar Ghattas. A unified discontinuous Petrov-Galerkin method and its analysis for Friedrichs' systems. *SIAM J. Numer. Anal.*, 51(4):1933–1958, 2013. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/DPGunifiedRevised.pdf.
32. Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves. *Inverse Problems*, 28(5):055001, 2012. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactI.pdf.
33. Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves. *Inverse Problems*, 28(5):055002, 2012. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactII.pdf.
34. Tan Bui-Thanh and Omar Ghattas. Analysis of the Hessian for inverse scattering problems. Part III: Inverse medium scattering of electromagnetic waves. Inverse Problems and Imaging, 2013. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/EM3Dmedium.pdf.
35. Bernardo Cockburn, Mitchell Luskin, Chi-Wang Shu, and Endre Süli. Post-processing of galerkin methods for hyperbolic problems. In Bernardo Cockburn, George E. Karniadakis, and Chi-Wang Shu, editors, *Discontinuous Galerkin Methods*, pages 291–300, Berlin, Heidelberg, 2000. Springer Berlin Heidelberg.
36. David Colton and Rainer Kress. *Integral equation methods in scattering theory*. John Wiley & Sons, 1983.
37. David Colton and Rainer Kress. *Inverse Acoustic and Electromagnetic Scattering*. Applied Mathematical Sciences, Vol. 93. Springer-Verlag, Berlin, Heidelberg, New-York, Tokyo, second edition, 1998.
38. S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White. MCMC methods for functions: modifying old algorithms to make them faster. *Statistical Science*, 28(3):424–446, 2013.
39. D. Cruz-Uribe and C. J. Neugebauer. An elementary proof of error estimates for the trapezoidal rule. *Mathematics Magazine*, 76(4):303–306, 2003.

40. Felipe Cucker and Ding Xuan Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, 2007.
41. D. J. Daley and J. Gani. *Epidemic Modelling*. Cambridge University Press, February 1984.
42. F. N. David. *Biometrika*, 42(3/4):540–540, 1955.
43. Philip J Davis and Philip Rabinowitz. *Methods of numerical integration*. Courier Corporation, 2007.
44. C. A. Desoer and B. H. Whalen. A note on pseudoinverses. *Journal of the Society for Industrial and Applied Mathematics*, 11(2):442–447, 1963.
45. O. Diekmann, J. A. P. Heesterbeek, and M. G. Roberts. The construction of next-generation matrices for compartmental epidemic models. *Journal of The Royal Society Interface*, 7(47):873–885, November 2009.
46. L. Couchman Dilip Ghosh Roy. *Inverse Problems and Inverse Scattering of Plane Waves*. Elsevier, 2001.
47. Simon Dobson. *Epidemic modelling - Some notes, maths, and code*. Independent Publishing Network, July 2020.
48. S. Duane, A. D. Kennedy, B. Pendleton, and D. Roweth. Hybrid Monte Carlo. *Phys. Lett. B*, 195:216–222, 1987.
49. Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, September 1936.
50. C H Edwards. *Advanced calculus of several variables*. Dover Books on Mathematics. Dover Publications, Mineola, NY, April 1995.
51. Alexandre Ern and Jean-Luc Guermond. *Theory and Practice of Finite Elements*, volume 159 of *Applied Mathematical Sciences*. Spinger-Verlag, 2004.
52. Alexandre Ern and Jean-Luc Guermond. Discontinuous Galerkin methods for Friedrichs' systems. Part I. General theory. *SIAM J. Numer. Anal.*, 44(2):753–778, 2006.
53. Alexandre Ern and Jean-Luc Guermond. *Theory and practice of finite elements*. Applied mathematical sciences. Springer, New York, NY, November 2010.
54. Alexandre Ern, Jean-Luc Guermond, and Gilbert Caplain. An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems. *Communications in partial differential equations*, 32(2):317–341, 2007.
55. Lawerence C. Evans. *Partial Differential Equations*. American Mathematical Society, Providence, RI, 1998.
56. Krzysztof J. Fidkowski. *A Simplex Cut-Cell Adaptive Method for High–order Discretizations of the Compressible Navier-Stokes Equations*. PhD thesis, Massachusetts Institute of Technology, 2007.
57. Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, March 1956.
58. Kurt O. Friedrichs. Symmetric positive linear differential equations. *Communications on pure and applied mathematics*, XI:333–418, 1958.
59. Karl Pearson F.R.S. Liii. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 1901.
60. Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980.
61. Matthias Gerdts. *Optimal Control of ODEs and DAEs*. De Gruyter, Berlin, Boston, 2012.
62. Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nyström method, and use of kernels in machine learning: Tutorial and survey, 2021.
63. D. Gilbarg and N.S. Trudinger. *Elliptic Partial Differential Equations of Second Order*. Springer-Verlag, Berlin, second (Classics in Mathematics) edition, 2001.

64. Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.
65. J. Hadamard. Sur les problèmes aux dérivés partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52, 1902.
66. Brian C. Hall. *Lie Groups, Lie Algebras, and Representations*. Springer International Publishing, 2015.
67. Takemitsu Hasegawa and Avram Sidi. An automatic integration procedure for infinite range integrals involving oscillatory kernels. *Numerical Algorithms*, 13:1–19, 1996.
68. Takemitsu Hasegawa and Hiroshi Sugiura. An automatic quadrature method for semi-infinite integrals of exponentially decaying functions and its matlab code. *Journal of Computational and Applied Mathematics*, 437:115450, 2024.
69. W. Keith Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.
70. Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.
71. J.M Heffernan, R.J Smith, and L.M Wahl. Perspectives on the basic reproductive ratio. *Journal of The Royal Society Interface*, 2(4):281–293, June 2005.
72. Herbert W. Hethcote. The mathematics of infectious diseases. *SIAM Review*, 42(4):599–653, January 2000.
73. R. A. Horn and C. A. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, Cambridge, London, New York, 1991.
74. Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3/4):321–377, 1936.
75. D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, March 1968.
76. Amy Hurford, Daniel Cownden, and Troy Day. Next-generation tools for evolutionary invasion analyses. *Journal of The Royal Society Interface*, 7(45):561–571, December 2009.
77. David James and Cynthia Botteron. Understanding singular vectors. *The College Mathematics Journal*, 44(3):220–226, 2013.
78. Max Jensen. *Discontinuous Galerkin methods for Friedrichs' systems with irregular solutions*. PhD thesis, University of Oxford, 2004.
79. Richard Arnold Johnson and Dean W. Wichern. *Applied multivariate statistical analysis*. Prentice Hall, Upper Saddle River, NJ, 5. ed edition, 2002.
80. Jari Kaipio and Erkki Somersalo. *Statistical and Computational Inverse Problems*, volume 160 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2005.
81. Kari Karhunen. Zur spektraltheorie stochastischer prozesse. 1946.
82. Tosio Kato. *Perturbation theory for linear operators; 2nd ed.* Grundlehren der mathematischen Wissenschaften : a series of comprehensive studies in mathematics. Springer, Berlin, 1976.
83. J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, September 1952.
84. Morris Kline. *Mathematical thought from ancient to modern times*. Oxford University Press, New York, NY, September 1972.
85. Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.
86. S. Kung. A new identificationand model reductionalgorithmvia singular value decompositions. In *Proc. 12th Ann. Asilomar Conf. Circuits, Syst. Comput.*, 1978.
87. Andrew J Kurdila and Michael Zabarankin. *Convex Functional Analysis*. Systems & Control: Foundations & Applications. Birkhauser Verlag AG, Basel, Switzerland, 2005 edition, May 2005.
88. Serge Lang. *Fundamentals of differential geometry*. Graduate Texts in Mathematics. Springer, New York, NY, 1 edition, September 2001.

89. A. Laub, M. Heath, C. Paige, and R. Ward. Computation of system balancing transformations and other applications of simultaneous diagonalization algorithms. *IEEE Transactions on Automatic Control*, 32(2):115–122, 1987.
90. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.
91. Glenn Ledder. *Mathematical Modeling for Epidemiology and Ecology*. Springer International Publishing, 2023.
92. Günter Leugering, Sebastian Engell, Andreas Griewank, Michael Hinze, Rolf Rannacher, Volker Schulz, Michael Ulbrich, and Stefan Ulbrich, editors. *Constrained Optimization and Optimal Control for Partial Differential Equations*. Springer Basel, October 2011.
93. Seppo Linnainmaa. The representation of the cumulative rounding error of an algorithm as a taylor expansion of the local rounding errors. *Master's Thesis (in Finnish), University of Helsinki*, pages 6–7, 1970.
94. Seppo Linnainmaa. Taylor expansion of the accumulated rounding error. *BIT*, 16(2):146–160, June 1976.
95. Hartmut Logemann and Eugene P. Ryan. *Ordinary Differential Equations Analysis, Qualitative Theory and Control*. Springer undergraduate mathematics series. Springer, London, 2014.
96. Juan Carlos De los Reyes. *Numerical PDE-Constrained Optimization*. Springer International Publishing, 2015.
97. D. G. Luenberger. *Optimization by Vector Space Methods*. John Wiley and Sons, New York, 1969.
98. Jonathan H Manton and Pierre-Olivier Amblard. A primer on reproducing kernel hilbert spaces. *Foundations and Trends® in Signal Processing*, 8(1–2):1–126, 2015.
99. Maia Martcheva. *An Introduction to Mathematical Epidemiology*. Springer US, 2015.
100. Carla D. Martin and Mason A. Porter. The extraordinary svd. *The American Mathematical Monthly*, 119(10):838–851, 2012.
101. James Martin, Lucas C. Wilcox, Carsten Burstedde, and Omar Ghattas. A stochastic Newton MCMC method for large-scale statistical inverse problems with application to seismic inversion. *SIAM Journal on Scientific Computing*, 34(3):A1460–A1487, 2012.
102. William McLean. *Strongly Elliptic Systems and Boundary Integral Equations*. Cambidge University Press, 2000.
103. Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.
104. Cleve Moler and Charles Van Loan. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. *SIAM Review*, 45(1):3–49, January 2003.
105. B. Moore. Principal component analysis in linear systems: Controllability, observability, and model reduction. *IEEE Transactions on Automatic Control*, 26(1):17–32, 1981.
106. E. H. Moore. On the reciprocal of the general algebraic matrix (abstract). *Bulletin of the AMS*, 26:394–395, 1920.
107. Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, ICML'10, page 807–814, Madison, WI, USA, 2010.
108. A. H. Nayfeh and O. N. Ashour. Acoustic receptivity of a boundary layer to Tollmien–Schlichting waves resulting from a finite-height hump at finite Reynolds numbers. *Physics of Fluids*, 6(11):3705–3716, 1994.
109. A.W. Naylor and G.R. Sell. *Linear Operator Theory in Engineering and Science*. Applied Mathematical Sciences. Springer New York, 1982.
110. R. M. Neal. *Handbook of Markov Chain Monte Carlo*, chapter MCMC using Hamiltonian dynamics. Chapman & Hall / CRC Press, 2010.

111. Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer Verlag, Berlin, Heidelberg, New York, second edition, 2006.
112. J. Tinsley Oden and Leszek F. Demkowicz. *Applied functional analysis*. CRC Press, 2010.
113. Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
114. Vern I. Paulsen and Mrinal Raghupathi. *An Introduction to the Theory of Reproducing Kernel Hilbert Spaces*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2016.
115. R. Penrose. A generalized inverse for matrices. *Mathematical Proceedings of the Cambridge Philosophical Society*, 51(3):406–413, 1955.
116. Antoine Perasso. An introduction to the basic reproduction number in mathematical epidemiology. *ESAIM: Proceedings and Surveys*, 62:123–138, 2018.
117. Niles A. Pierce and Michael B. Giles. Adjoint recovery of superconvergent functionals from pde approximations. *SIAM Review*, 42(2):247–264, 2000.
118. Anil V. Rao. Riccati dichotomic basis method for solving hypersensitive optimal control problems. *Journal of Guidance, Control, and Dynamics*, 26(1):185–189, 2003.
119. A.V. Rao and K.D. Mease. Dichotomic basis approach to solving hyper-sensitive optimal control problems. *Automatica*, 35(4):633–642, 1999.
120. J Tilak Ratnanather, Jung H Kim, Sirong Zhang, Anthony MJ Davis, and Stephen K Lucas. Algorithm 935: Iipbf, a matlab toolbox for infinite integral of products of two bessel functions. *ACM Transactions on Mathematical Software (TOMS)*, 40(2):1–12, 2014.
121. Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, September 1951.
122. Christian P. Robert and George Casella. *Monte Carlo Statistical Methods (Springer Texts in Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2005.
123. Kelly Jabbusch Robert Graff and David James. Intuitive interpretations of svd vectors. *The College Mathematics Journal*, 54(3):200–211, 2023.
124. Gareth O. Roberts and Jeffrey S. Rosenthal. Optimal scaling of discrete approximations to Langevin diffusions. *J. R. Statist. Soc. B*, 60:255–268, 1997.
125. Gareth O. Roberts and Richard L. Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.
126. Halsey Royden and Patrick Fitzpatrick. *Real Analysis (Classic Version)*. Pearson, Upper Saddle River, NJ, 4 edition, February 2017.
127. W. Rudin. *Functional Analysis*. McGraw-Hill, New York, St. Louis, San Francisco,..., 1973.
128. Jürgen Schmidhuber. Deep learning in neural networks: An overview. *Neural Networks*, 61:85–117, January 2015.
129. (1915-2002) Schwartz, Laurent. *Théorie des distributions*. Publications de línstitut de mathematique de lúniversite de strasbourg,9; 10. Hermann, Paris, nouv. édition entiérement corrigée, refondue et augmentée edition, DL 1966.
130. A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. Society for Industrial and Applied Mathematics, 2009.
131. R. E. Showalter. *Hilbert Space Methods for Partial Differential Equations*. Pitman, London, San Francisco, Melbourne, 1977.
132. R.E. Showalter. *Hilbert Space Methods in Partial Differential Equations*. Dover books on mathematics. Dover Publications, 2011.

133. Lawrence Sirovich. Turbulence and the dynamics of coherent structures. i. coherent structures. *Quarterly of Applied Mathematics*, 45(3):561–571, 1987.
134. G. W. Stewart. On the early history of the singular value decomposition. *SIAM Review*, 35(4):551–566, 1993.
135. Stephen M. Stigler. Gauss and the Invention of Least Squares. *The Annals of Statistics*, 9(3):465 – 474, 1981.
136. Gilbert Strang. The fundamental theorem of linear algebra. *Am. Math. Mon.*, 100(9):848, November 1993.
137. Aleksandr N Tikhonov, A Goncharsky, V V Stepanov, and A G Yagola. *Numerical methods for the solution of ill-posed problems*. Mathematics and Its Applications. Springer, Dordrecht, Netherlands, 1995 edition, June 1995.
138. Andrey N Tikhonov and Vasiliy Y Arsenin. *Solutions of ill-posed problems*. Scripta series in mathematics. John Wiley & Sons, Nashville, TN, August 1977.
139. P. van den Driessche and James Watmough. Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, 180(1-2):29–48, November 2002.
140. Wikipedia contributors. Singular value decomposition — Wikipedia, the free encyclopedia. https://en.wikipedia.org/w/index.php?title=Singular_value_decomposition&oldid=1185343144, 2023. [Online; accessed 22-November-2023].
141. Kôsaku Yosida. *Functional Analysis*. Springer Berlin Heidelberg, 1995.
142. E Zeidler. *Nonlinear functional analysis and its applications*. Springer, New York, NY, 1985 edition, November 1984.
143. Eberhard Zeidler. *Applied Functional Analysis*. Springer New York, 1995.
144. O. C. Zienkiewicz and J. Z. Zhu. A simple error estimator and adaptive procedure for practical engineerng analysis. *International Journal for Numerical Methods in Engineering*, 24(2):337–357, 1987.

Index