Oden Institute REPORT 23-04

July 2023

Adjoint and Its roles in Sciences, Engineering, and Mathematics: A Tutorial

by

Tan Bui-Thanh

Oden Institute for Computational Engineering and Sciences
The University of Texas at Austin
Austin, Texas 78712

Adjoint and Its roles in Sciences, Engineering, and Mathematics: A Tutorial

Tan Bui-Thanh*

Abstract. This paper synergizes the roles of adjoint in various disciplines of mathematics, sciences, and engineering. Though the materials developed and presented are not new—as each or some could be found in (or inferred from) publications in different fields—we believe this is the first effort to systematically unify these materials on the same mathematical footing starting from the basic definitions. We aim to provide a unified perspective and understanding of adjoint applications. As a result, this work could give broader views and better insights into the application of adjoint beyond a single community. By rigorously establishing general results and then developing materials specific to each application, we bring forth details on how abstract concepts/definitions can be translated into particular applications and the connections among them. This paper is written as an interdisciplinary tutorial on adjoint with discussions and with many examples from different fields including linear algebra (e.g. eigendecomposition and the singular value decomposition), ordinary differential equations (asymptotic stability of an epidemic model), partial differential equations (well-posedness of elliptic, hyperbolic, and Friedrichs' types), neural networks (backpropagation of feed-forward deep neural networks), least squares and inverse problems (with Tikhonov regularization), and PDE-constrained optimization. The exposition covers results and applications in both finite-dimensional and infinite-dimensional Hilbert spaces.

Key words. Adjoint, optimization, backpropagation, eigenvalue problem, singular value decomposition, asymptotic stability, wellposedness, least squares, PDE-constrained optimization, reproduction number.

1. Introduction. Adjoint is ubiquitous in mathematics. According to [24], the history of adjoint can be traced back to as far as Lagrange in 1766 [48], in a memoir extending the letter that he wrote to D'Alembert in January 1765 discussing his new method of solving nth-order differential equations. Also according to [24], the term "adjoint equation", first used to call the corresponding equation developed from Lagrange memoir [48], is due to Fuchs [34]. The adjoint operator, as reported in [53], was introduced by Riesz in his seminal paper[1] [65] to study the inverse of linear operators.

Since then adjoint has been pervasive in vast literature across mathematics, engineering, and sciences disciplines. This is not surprising as the adjoint has many appealing features including i) the adjoint operator typically possesses nicer properties than the original operator (e.g. the adjoint of a densely defined linear operator is a closed operator though the original operator may not), and ii) the adjoint equation is always linear even when the original equation is not. Though a comprehensive survey on adjoint accounting for its application in many disciplines/fields (and their sub-disciplines) could be desirable to appreciate the crucial role that adjoint plays, it is perhaps an impossible task.[2]

Our objective is to provide a window into the adjoint and its crucial role in certain sub-

---

*Department of Aerospace Engineering and Engineering Mechanics, The Oden Institute for Computational Engineering and Sciences, the University of Texas at Austin, Texas (tanbui@oden.utexas.edu, https://users.oden.utexas.edu/~tanbui/).

[1]As also argued in [53], in his seminal paper [65] Riesz established *functional analysis* as a new mathematical discipline.

[2]Or more precisely, it is rather the task for a book than for a paper.

sets of computational science, engineering, and mathematics. The exposition is necessarily personal and biased based on topics that we are familiar with. The materials developed and presented are not new, as each or some could be found in (or inferred from) publications in different fields. Our objective is to systematically unify these materials on the same mathematical footing starting from the basic definitions. This expectantly provides a more unified perspective on the usefulness of adjoint in variety of applications. As a result, this work could give broader views and better insights into applications of adjoint beyond one 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. This paper is written as a tutorial on adjoint with many examples presented with discussions. Though we strike for a self-contained expository, it is necessary for us to state a few results without proof to keep the length of the paper manageable and to focus on the adjoint and its roles.

The paper is structured as follows. In section 2 we introduce various notations, definitions, and some examples. The paper is then developed into two parts: Part I in finite dimensions (section 3) and Part II in infinite dimensions (section 4). In order to keep the exposition succinct, definitions and results that are valid for both cases are presented/proved once and when that happens we will explicitly state so. Most of our developments start from abstract operator settings and then reduce to standard finite-dimensional settings in $\mathbb{R}^n$ as a special case. In some cases, such as optimization, the order is reversed as we believe it is more natural that way. Each section of the paper is equipped with examples on which we show how to apply the preceding abstract theoretical results. We make an effort to include practical examples from different fields including linear algebra (e.g. eigendecomposition and the singular value decomposition), ordinary differential equations (an epidemic model), partial differential equations (of elliptic, hyperbolic, and Friedrichs' types), neural networks (feed-forward deep neural networks), least squares and inverse problems (with Tikhonov regularization), PDE-constrained optimization, etc. Due to the interdisciplinary nature of the paper, we do not attempt to provide an exhaustive list of references but a few for each section to keep the references at a manageable length.

For section 3, we begin with the celebrated Riesz representation theorem that is then used to prove the existence of the adjoints of continuous linear operators. It is followed by the closed range theorem that will be useful in many places later. Built upon these basic materials, we shall develop several applications of adjoint. The first application is in subsection 3.1 in which we highlight the role of adjoint in assessing the solvability of linear operator equations before solving them. Perhaps one of the most important applications of adjoint is in the study of eigenvalue problems and this is the main focus of subsection 3.2. The main result for this section is the spectral decomposition of self-adjoint operators in finite dimensions. Another important one is the tight relationship between the orthogonality of a projection and its self-adjointness, which immediately leads to a generalized Pythagorean theorem. In subsection 3.3, we start with the classical projection theorem and then deploy it together with the closed range theorem to find the necessary and sufficient condition for the optimality of an abstract linear least squares problem. The next important application is the singular value decomposition (SVD) in subsection 3.4, in which we employ the spectral decomposition in subsection 3.2 to establish an SVD decomposition for general linear operator in finite dimensions. This 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. We then discuss the equivalence of the SVD of an abstract linear operator and the SVD of its matrix representation.

Our next application of interest for section 3 is optimization with equality constraints. This is the main topic of subsection 3.5 in which we expose at length the role of adjoint in optimization theory that is valid for both finite and infinite dimensions. This is accomplished by working with Fréchet derivative and its Riesz representation counterpart as the gradient. Though can be further developed (e.g. to second-order optimality conditions) our focus for this expository is on the first-order optimality condition. We recall an implicit function theorem and use it to prove an abstract inverse function theorem, which is then deployed to derive the first-order optimality condition for abstract optimization problems with equality constraints. The important role of adjoint comes into the picture when we prove an abstract Lagrangian multiplier theorem using the closed range theorem. The importance of adjoint is further amplified when we study constrained optimization problems with separable structure. Here, adjoint facilitates an efficient gradient-based optimization algorithm in unconstrained reduced subspace while ensuring the feasibility of the contraints at all times. Though discussed in the separate subsection 3.6, we show that when applying this reduced space approach for the optimization problem arisen from training deep neural networks (DNNs), we recover the backpropagation algorithm that has been the workhorse in training DNNs. Viewing backpropagation from the reduced space approach provides further insights into the algorithm as we shall argue. The last finite dimensional application that we present in subsection 3.7 is the stability of autonomous ordinary differential equations (ODEs). The main goal of this section is to exhibit the vital role of adjoint in establishing the necessary and sufficient conditions for the asymptotic stability (in the sense of Lyapunov) of ODEs' equilibiria.

For the infinite dimensional settings in section 4, we start with a more general adjoint definition for densely defined linear operators. This is useful for all subsections except subsection 4.1. The first application that we study is the illposedness (in Hadamard's sense) nature of inverting compact operators in subsection 4.1. To that end, we extend the spectral theorem in subsection 3.2 and SVD decomposition theorem in subsection 3.4 to the Hilbert-Schmidt theorem and a general SVD theorem for compact (linear) operators in infinite dimensions. The main result is a Picard theorem which states the necessary and sufficient conditions under which the task of inverting a compact operator is solvable. This is then employed to show that inverting a compact operator violates at least the stability condition of well-posedness. We then deploy the Riesz-Fredholm theory to show how Tikhonov regularization can restore the well-posedness at the expense of getting a nearby solution. In subsection 4.2, we discuss the role of adjoint in establishing the well-posedness of abstract linear operator equations with application to partial differential equations (PDEs). Two main results developed in this section are the Banach-Nečas-Babuška theorem and the Lax-Milgram lemma. We then switch to subsection 4.3 to study Sturm-Liouville eigenvalue problem and generalized Fourier series in $\mathbb{L}^2$. Our exposition is for closed linear operators. Using the Hilbert-Schmidt theorem in subsection 4.1 and the Lax-Milgram lemma in subsection 4.2 we establish a spectral decomposition theorems for quite general linear operators, and then apply them to Sturm-Liouville eigenvalue problems to derive Fourier series and its generalization. The last application of

interest is PDE-constrained optimization and this is the main topic of subsection 4.4. Here we show how to rigorously translate the abstract Lagrangian multiplier theorem in subsection 3.5 to derive the adjoint equation and the reduced gradient for prototype elliptic and hyperbolic PDEs. We show that the differential operators of adjoint equations are indeed the adjoint operators that we derive at the beginning of section 4.

2. Notations. In this paper, boldface lower case letters such as $\boldsymbol{u}$ are reserved for vector-valued functions in $\mathbb{R}^n$, for some integer n. Calligraphic uppercase letters such as $\mathcal{A}$ 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 [2, 60]. Lowercase letters are for scalar-valued functions. *We also use lowercase letters for results that are valid for both finite and infinite dimensional settings* and boldface uppercase letters are for bases of vector spaces. 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 $\|u\|_{\mathbb{X}} = \sqrt{(u,u)_{\mathbb{X}}}$. All spaces are either complex or real. For the former, the inner product is conjugate symmetric, i.e., $(u,v)_{\mathbb{X}} = \overline{(v,u)_{\mathbb{X}}}$, where the overline denotes the complex conjugate. 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. We shall frequently identify the dual of any Hilbert space with itself. We define $\mathfrak{L}(\mathbb{X}, \mathbb{Y})$ as the space of all linear operators from $\mathbb{X}$ to $\mathbb{Y}$, $\mathfrak{B}(\mathbb{X}, \mathbb{Y})$ as the space of all bounded linear operators from $\mathbb{X}$ to $\mathbb{Y}$, and $\mathfrak{C}(\mathbb{X}, \mathbb{Y})$ as the space of all continuous mapping from $\mathbb{X}$ into $\mathbb{Y}$. By $\mathfrak{C}^n(\mathbb{X})$ and $\mathfrak{C}_0^\infty(\mathbb{X})$ 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). 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".

Definition 2.1 (Linear transformation/map/operator). *Consider two inner product vector spaces, $\mathbb{X}$, $(\cdot, \cdot)_{\mathbb{X}}$ and $\mathbb{Y}$, $(\cdot, \cdot)_{\mathbb{Y}}$. Suppose $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ satisfies the following*

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

where $\alpha, \beta \in \mathbb{C}$, and $u, w \in \mathbb{X}$. Then, we call[3] $\mathscr{A}$ a linear transformation or map, or an operator from $\mathbb{X}$ to $\mathbb{Y}$.

Convention 2.2.

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

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

3. The range of $\mathscr{A}$ is defined as

$$\mathsf{R}(\mathscr{A}) := \{\mathscr{A}(u) : u \in \mathsf{D}(\mathscr{A})\} = \{y \in \mathbb{Y} : \exists u \in \mathsf{D}(\mathscr{A}) \text{ and } y = \mathscr{A}(u)\} \subset \mathbb{Y}.$$

---

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

4. The kernel of $\mathscr{A}$ (or the null space of $\mathscr{A}$) is defined as

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

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

Example 2.3. Let us consider $\mathbb{X} = \mathbb{R}^n, \mathbb{Y} = \mathbb{R}^m$, and let $\mathcal{A} : \mathbb{X} \mapsto \mathbb{Y}$ be an $\mathbb{R}^{m \times n}$ matrix. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{X}$, $\alpha, \beta \in \mathbb{F}$, we have

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

Thus, any matrix is a linear operator.

Example 2.4. Let $\mathscr{A} : \mathbb{X} = \mathbb{L}^2(0,1) \to \mathbb{Y} = \mathbb{R}$ 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 $\alpha, \beta \in \mathbb{F}$:

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

and thus integrals are linear operators.

Example 2.5. Consider $\mathbb{X} = \mathbb{Y} = \mathbb{L}^2(0,1)$ and $\mathscr{A} : D(\mathscr{A}) \subset \mathbb{X} \to \mathbb{Y}$ such that

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

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

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

thus differentiation is a linear operator.

3. Part I: Adjoint operators in finite dimensional Hilbert spaces. In this part, unless otherwise stated, we assume that $\mathbb{X}$ and $\mathbb{Y}$ are finite dimensional vector spaces, i.e. $\dim \mathbb{X} = n < \infty$ and $\dim \mathbb{Y} = m < \infty$, where $\dim$ denotes the dimension. Recall that if $\mathscr{A} \in \mathfrak{L}(\mathbb{X}, \mathbb{Y})$ and $\dim \mathbb{X} < \infty$, then $\mathscr{A} \in \mathscr{B}(\mathbb{X}, \mathbb{Y})$. Let $\boldsymbol{E} = \{e_1, \dots, e_n\}$ and $\boldsymbol{G} = \{g_1, \dots, g_m\}$ be orthonormal[4] 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 $\mathcal{A}^{\boldsymbol{EG}}$. 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 $\mathcal{A}$ as $\mathcal{A}(i, j)$. 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.

This section 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 subsection 3.1 on the solvability of linear operator equations before solving them. The role of adjoint in the study of eigenvalue problems is given in subsection 3.2. In subsection 3.3, 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 subsection 3.4. 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 subsection 3.5 in which we expose at length the role of adjoint in optimization theory that is valid for both finite and infinite dimensions. This follows by showing that a reduced spaced approach using adjoint reduces to backpropagation of deep neural networks in subsection 3.6. The last finite dimensional application that we present in subsection 3.7 is role of adjoint in establishing the stability of autonomous ordinary differential equations (ODEs).

Theorem 3.1 (Riesz representation theorem). *Let $\mathscr{L}$ be a bounded linear functional on a Hilbert space $\mathbb{X}$. There exists a unique $u \in \mathbb{X}$ such that*

$$\mathscr{L}(v) = (u, v)_{\mathbb{X}}, \quad \forall v \in \mathbb{X}.$$

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

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., [63, 6, 16, 58, 72]). We provide a short and intuitive proof for finite dimensions. We prove the result for $\mathbb{F} = \mathfrak{C}$ as the case $\mathbb{F} = \mathbb{R}$ is analogous. Let $\boldsymbol{E} = \{e_1, \dots, e_n\}$ be an orthonormal basis for $\mathbb{X}$. Let $\boldsymbol{u}$ be the representation of u in $\boldsymbol{E}$. we have

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

---

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

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

Definition 3.2 (Adjoint operator). *Let* $\mathscr{A} \in \mathfrak{B}(\mathbb{X}, \mathbb{Y})$. *We say that* $\mathscr{A}^* : \mathbb{Y} \to \mathbb{X}$ *is the adjoint of A iff*

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

Proposition 3.3. *Let* $\mathscr{A} \in \mathfrak{B}(\mathbb{X}, \mathbb{Y})$. *Then* $\mathscr{A}^*$ *exists and is unique. Furthermore, it is linear with* $\|\mathscr{A}^*\| = \|\mathscr{A}\|$, *where the operator norm is defined as usual, e.g.,*

$$\|\mathscr{A}\| := \sup_{u \in \mathbb{X}} \frac{\|\mathscr{A}u\|_{\mathbb{Y}}}{\|u\|_{\mathbb{X}}} = \sup_{\|u\|_{\mathbb{X}}=1} \|\mathscr{A}u\|_{\mathbb{Y}}.$$

Proof. To see the existence and linearity, we note that owing to the continuity of $\mathscr{A}$ and inner product, $(\mathscr{A}u, v)_{\mathbb{Y}}$ is continuous in u. By the Riesz representation Theorem 3.1, there exists a unique $\ell \in \mathbb{X}$ depending on $\mathscr{A}$ and v such that

$$(u, \ell(\mathscr{A}, v))_{\mathbb{X}} = (\mathscr{A}u, v)_{\mathbb{Y}}, \quad \forall u \in \mathbb{X},$$

which implies that $\ell(\mathscr{A}, v)$ is linear in v. Defining $\mathscr{A}^*v := \ell(\mathscr{A}, v)$ shows the $\mathscr{A}^*$ exists and linear. We next show that $\mathscr{A}^*$ is continuous (bounded) and $\|\mathscr{A}^*\| = \|\mathscr{A}\|$. We have

$$\|\mathscr{A}^*v\|_{\mathbb{X}}^2 = (\mathscr{A}^*v, \mathscr{A}^*v)_{\mathbb{X}} = (\mathscr{A}\mathscr{A}^*v, v)_{\mathbb{Y}} \leqslant \|\mathscr{A}\| \|\mathscr{A}^*v\|_{\mathbb{X}} \|v\|_{\mathbb{Y}},$$

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

Remark 3.4. Note that Theorem 3.1, Definition 3.2, and Proposition 3.3 are also valid for infinite-dimensional settings.

Example 3.5. Let $\mathbb{U} = \mathbb{R}^n$ and $\mathbb{V} = \mathbb{R}^n$ be respectively endowed with the inner products $(.,.)_{\mathbb{R}^n}$ and an $\mathcal{M}$-weighted inner product $(.,.)_{\mathbb{R}^n, \mathcal{M}}$ where $(\boldsymbol{v}, \boldsymbol{w})_{\mathbb{R}^n, \mathcal{M}} := \sum_{i,j} \boldsymbol{v}(i) \mathcal{M}(i, j) \boldsymbol{w}(j) := \boldsymbol{v}^T \mathcal{M} \boldsymbol{w}, \forall \boldsymbol{v}, \boldsymbol{w} \in \mathbb{V}$, and $\mathcal{M}$ is a symmetric and positive definite matrix. We need to find the adjoint operator $\mathcal{A}^*$ of a matrix $\mathcal{A} : \mathbb{U}, (.,.)_{\mathbb{R}^n} \to \mathbb{V}, (.,.)_{\mathbb{R}^n, \mathcal{M}}$. We have

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

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

Example 3.6. Now, let us consider $\mathscr{A} : \mathbb{U} = \text{Span}\{1, \ x, \ x^2\} \subset \mathbb{X} = \mathbb{L}^2(-1, 1) \to \mathbb{R}^2$ such that the map A is defined as

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

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

$$(\mathscr{A}u,\ \boldsymbol{v})_{\mathbb{R}^2} = \boldsymbol{v}_1 \int_{-1}^{1} u(x)\ dx + \boldsymbol{v}_2 \int_{-1}^{1} (2x+1)\ u(x)\ dx$$

$$= \int_{-1}^{1} u(x)\left[\boldsymbol{v}_1 + (2x+1)\boldsymbol{v}_2\right] dx = \int_{-1}^{1} u(x)\left[1,\ (2x+1)\right]\boldsymbol{v}\ dx.$$

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

Example 3.7. Let us consider $\mathscr{A} : \mathbb{U} = \mathrm{Span}\{1,\ x,\ x^2\} \subset \mathbb{X} = \mathbb{L}^2(-1,1) \to \mathbb{R}^2$ such that the map A is defined as

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

where $x_1, x_2 \in (-1,1)$ and $x_1 \neq x_2$. 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

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

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

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

and thus

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

by testing (3.1) with $u \in \{1, x, x^2\}$.

Example 3.8. 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

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

which by definition gives

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

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

(3.2)
$$\int_0^1 u(x)\overline{\delta(x-1)}\,dx = u(1), \quad \forall u \in \mathbb{U}.$$

Clearly, we can find $\delta(x-1)$ as the unique polynomial of degree at most n by testing (3.2) with $u \in \{1, x, \ldots, x^n\}$.

Proposition 3.9. *Let $\boldsymbol{E}$ and $\boldsymbol{G}$ be orthonormal bases of $\mathbb{X}$ and $\mathbb{Y}$, respectively, and $\dim\mathbb{X} = n$ and $\dim\mathbb{Y} = m$. Let $\mathcal{A}$ and $\mathcal{B}$ be the matrix representations of $\mathscr{A}$ and $\mathscr{A}^*$ with respect to the bases $\boldsymbol{E}$ and $\boldsymbol{G}$. Then*

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

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

Proof. By the definition of adjoint and matrix representation, for any $u \in \mathbb{X}$ and $v \in \mathbb{Y}$ we have

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

which concludes the proof. ∎

The following Definition 3.10, Definition 3.11, Corollary 3.12, Theorem 3.13, and Corollary 3.14 are valid for both finite and infinite dimensions.

Definition 3.10 (Orthogonal complement). *Let $\mathbb{S} \subset \mathbb{X}$, the orthogonal complement $\mathbb{S}^\perp$ of $\mathbb{S}$ is defined as*

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

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

Definition 3.11 (Closure). *Let $\mathbb{S} \in \mathbb{X}$. The closure $\overline{\mathbb{S}}$ of $\mathbb{S}$ is the smallest closed set containing $\mathbb{S}$.*

Note that we use the overline "$\overline{\cdot}$" to denote both complex conjugate and the closure, but it should be clear from the context.

Corollary 3.12. *There holds:* $\left(\mathbb{S}^\perp\right)^\perp = \overline{\mathbb{S}}$.

Proof. See [58, proposition 1 of chapter 3]. ∎

Next is an important theorem (see, e.g., [77, 16, 8]).

Theorem 3.13 (The closed range theorem). *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$. The following hold:*

- $[R(\mathscr{A})]^{\perp} = N(\mathscr{A}^*).$
- $\overline{R(\mathscr{A})} = [N(\mathscr{A}^*)]^{\perp}.$
- $[R(\mathscr{A}^*)]^{\perp} = N(\mathscr{A}).$
- $\overline{R(\mathscr{A}^*)} = [N(\mathscr{A})]^{\perp}.$[5]

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

  Proof. The second assertion is the direct consequence of the first assertion and Corollary 3.12. 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 $z \in N(\mathscr{A}^*)$ and $y \in R(A)$. Then $y = \mathscr{A}x$ for some $x \in \mathbb{X}$. We have

  $$(z, y)_{\mathbb{Y}} = (z, \mathscr{A}x)_{\mathbb{Y}} = (\mathscr{A}^* z, x)_{\mathbb{X}} = 0, \forall y \in R(A),$$

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

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

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

  Corollary 3.14. *There hold:*
- $\mathbb{X} = N(\mathscr{A}) \oplus \overline{R(\mathscr{A}^*)}$
- $\mathbb{Y} = N(\mathscr{A}^*) \oplus \overline{R(\mathscr{A})}$

  We note that for finite dimensional vector spaces $\mathbb{X}$ and $\mathbb{Y}$, $R(\mathscr{A})$, and hence $R(\mathscr{A}^*)$, is obviously closed, and thus Theorem 3.13 and Corollary 3.14 clearly hold. We will see that the proof of the closed range Theorem 3.13, hence the Corollary 3.14, for finite dimensional cases is trivial using the SVD decomposition in Theorem 3.33.

  3.1. Application of adjoint to the solvability of linear operator equations. In this section, we study the existence of a solution of the following linear operator equation

$$(3.3) \qquad \qquad \mathscr{A}u = f,$$

  where $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$.

  Lemma 3.15.
- **Existence.** *The linear equation* (3.3) *has a solution iff $y \in N(\mathscr{A}^*)^{\perp}$.*
- **Uniqueness.** *The solution of* (3.3) *is unique iff $N(\mathscr{A}) = \{\theta\}$.*
- *If $\dim \mathbb{X} = \dim \mathbb{Y}$, the uniqueness is equivalent to existence.*

  Proof. The existence is the direct consequence of Theorem 3.13. The uniqueness is clear.

---

[5]Since we consider only Hilbert spaces, which are reflexive, $\overline{R(\mathscr{A}^*)} = [N(\mathscr{A})]^{\perp}$ holds. In general, $\overline{R(\mathscr{A}^*)} \subset [N(\mathscr{A})]^{\perp}$: see [16, Corollary 2.18].

The proof of the third assertion is as follows. We have

$$\mathsf{N}(\mathscr{A}) = \{\theta\} \hspace{4cm} \text{Uniqueness}$$

$$\Updownarrow$$

$$\dim \mathsf{N}(\mathscr{A}) = 0$$

$$\Updownarrow \hspace{2cm} \dim \mathbb{X} = \dim \mathbb{Y} \text{ and } \dim \mathbb{X} = \dim \mathsf{N}(\mathscr{A}) + \dim \mathsf{R}(\mathscr{A})$$

$$\dim \mathbb{Y} = \dim \mathbb{X} = \dim \mathsf{R}(\mathscr{A})$$

$$\Updownarrow \hspace{5cm} \mathsf{R}(\mathscr{A}) \subseteq V$$

$$\mathsf{R}(\mathscr{A}) = \mathbb{Y} \hspace{3cm} \text{Existence for any } y \in \mathbb{Y} \quad \blacksquare$$

Remark 3.16. The existence condition can be simply $y \in \mathbb{R}(\mathscr{A})$. However, it is easier to work with $\mathsf{N}(\mathscr{A}^*)^{\perp}$ as it gives us equations (see Definition 3.10) to determine/characterize $\mathsf{N}(\mathscr{A}^*)^{\perp}$.

Example 3.17. Consider the operator $\mathscr{A}$ defined in Example 3.6, and we are interested in studying the existence of a solution[6] for $\mathscr{A}u = \boldsymbol{f}$. We start by recalling from Example 3.6 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

$$v_1 + (2x + 1)v_2 = 0 \quad \forall x,$$

which implies

$$v_1 = v_2 = 0.$$

Thus,

$$\mathsf{N}(\mathscr{A}^*) = \{\theta\}.$$

Since θ is orthogonal to any $\boldsymbol{f}$, we conclude that $\mathscr{A}u = \boldsymbol{f}$ always has a solution.

Example 3.18. Let $\mathcal{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}$, and $\boldsymbol{f} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The question is if there is a solution to the equation $\mathcal{A}\boldsymbol{u} = \boldsymbol{f}$. Consider $\mathbb{U} = \mathbb{V} = \mathbb{R}^2$ with the standard Euclidean inner product $(\cdot, \cdot)_{\mathbb{R}^2}$. We know from Example 3.5 that the adjoint $\mathcal{A}^*$ is given by

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

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

$$\mathcal{A}^* \mathbf{z} = \theta \implies z_1 + z_2 = 0,$$

---

[6]Note that this is an operator setting for the problem of fitting a quadratic polynomial $u(x)$ with two pieces of information about $u(x)$.

which yields $N(\mathcal{A}^*) = \left\{ \begin{bmatrix} \alpha \\ -\alpha \end{bmatrix} : \quad \forall \alpha \in \mathbb{R} \right\}$. However, for any $\boldsymbol{z} \in N(\mathcal{A}^*)$ 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, $\boldsymbol{f} \not\perp N(\mathcal{A}^*)$. Thus, $\mathcal{A}\boldsymbol{u} = \boldsymbol{f}$ does not have a solution.

3.2. Application of adjoint to eigenvalue problems.

Definition 3.19 (Eigenvalue problem). *Let $A : \mathbb{X} \to \mathbb{X}$ be a linear operator.*

(3.4) $\mathscr{A}u = \lambda u \qquad \forall \lambda \in \mathbb{C}, \ u \in \mathbb{X}$

is called an eigenvalue problem if there exists a nontrivial pair (λ, x) (x is not a zero vector/function but λ could be zero). In particular:
- *λ is called an eigenvalue.*
- *x is called an eigenfunction, associated with the eigenvalue λ, of $\mathscr{A}$. If $\mathbb{X}$ is a finite-dimensional space, i.e., $\mathbb{X} = \mathbb{R}^n$, x is typically called eigenvector.*

Definition 3.20 (Self-adjoint operator). *If $\mathscr{A}^* = \mathscr{A}$, then $\mathscr{A}$ is called self-adjoint.*

Lemma 3.21. *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 (λ, u) and (α, v) are two eigen-pairs and $\lambda \neq \alpha$ then $(u, v)_{\mathbb{X}} = 0$.*

Proof. For the first assertion, we have $\lambda (u, v)_{\mathbb{X}} = (\lambda u, u)_{\mathbb{X}} = (\mathscr{A}u, u)_{\mathbb{X}} = (u, \mathscr{A}u)_{\mathbb{X}} = (u, \lambda u)_{\mathbb{X}} = \overline{\lambda} (u, \lambda u)_{\mathbb{X}}$. Thus, $(\lambda - \overline{\lambda}) (u, \lambda u)_{\mathbb{X}} = 0$, 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 3.22. Note that Definition 3.19, Definition 3.20, and Lemma 3.21 hold for both finite and infinite dimensional cases.

Proposition 3.23. *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 $\mathcal{A}$ 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 (3.4) reads

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

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

Theorem 3.24. *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 3.23 implies that $\mathscr{A}$ has at least one eigenfunctions. Let $\mathbb{S}$ be the span of all eigenfunctions of $\mathscr{A}$. Owing to Lemma 3.21, it is sufficient to show that $\mathbb{S} = \mathbb{X}$. If $\mathbb{S}^{\perp} = \{\theta\}$, then clearly $\mathbb{S} = \mathbb{X}$. Now suppose $\mathbb{S}^{\perp} \neq \{\theta\}$. If $u \in \mathbb{S}^{\perp}$ and (λ, 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 $v \in \mathbb{S}$. Thus $\mathscr{A} : \mathbb{S}^{\perp} \to \mathbb{S}^{\perp}$. By Proposition 3.23, $\mathscr{A}$ has an eigenfunction w in $\mathbb{S}^{\perp}$, which means that $w \in \mathbb{S} \cap \mathbb{S}^{\perp}$, which in turn implies $w = \theta$, a contradiction. We conclude that $\mathbb{S}^{\perp} = \{\theta\}$ and this concludes the proof. ∎

Corollary 3.25 (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, $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \cdots \geqslant \lambda_n$ 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 determined by its eigenpairs.*

Proof. The first assertion is clear due to Lemma 3.21, Proposition 3.23, and Theorem 3.24. The second assertion is also obvious since

$$(3.5) \qquad\qquad x = \sum_{i=1}^{n} (x, u)_{\mathbb{X}} u_i,$$

due to Theorem 3.24. ∎

Example 3.26 (Eigen-decomposition of self-adjoint (Hermitian) matrices). Let $\mathcal{A} : \mathbb{F}^n \to \mathbb{F}^n$ be a self-adjoint matrix. Applying Corollary 3.25 we conclude that $\mathcal{A}$ has n real eigenvalues $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \cdots \geqslant \lambda_n$ and orthonormal eigenvectors $\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots \boldsymbol{u}_n$ such that $\mathcal{A}\boldsymbol{u}_i = \lambda_i \boldsymbol{u}_i$ and

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

where $U \in \mathbb{F}^{n \times n}$ is the eigenmatrix whose columns are eigenvectors of $\mathcal{A}$ and Λ is a diagonal matrix whose diagonals are the corresponding eigenvalues of $\mathcal{A}$. Thus, the standard eigendecomposition for self-adjoint matrices is a special case of Corollary 3.25.

We next discuss the relationship between self-adjointness and orthogonal projection.

Definition 3.27 (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 3.28. *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 3.13, but we omit the details for brevity. ∎

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

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

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 into $\mathbb{X}$. More generally, we can verify that

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

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

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

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

$$(3.6) \qquad \|x\|_{\mathbb{X}}^2 = \|(\mathscr{I} - \mathscr{P})\,x\|_{\mathbb{X}}^2 + \|\mathscr{P}x\|_{\mathbb{X}}^2 .$$

3.3. Application of adjoint to linear least squares problems. We start with the classical projection theorem that holds for both finite and infinite dimensional settings.

Theorem 3.29 (Projection theorem). *Let $\mathbb{S}$ be a subspace of a pre-Hilbert[7] space $\mathbb{Y}$. Let $y \in \mathbb{Y}$. Then, $u \in \mathbb{S}$ is the unique minimizer of $\inf_{w \in \mathbb{S}} \|w - y\|_{\mathbb{Y}}$ iff $(y - u) \perp \mathbb{S}$. The existence of the minimizer u is guaranteed if $\mathbb{Y}$ is Hilbert and $\mathbb{S}$ is closed.*

Proof. We follow closely the proof by contradiction in [58, Theorem 2]. Suppose there exists $v \in \mathbb{S}$ 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

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

contradicting the fact that u is a minimizer. Conversely, let $(y - u) \perp \mathbb{S}$, by the Pythagorean identity (3.6), for any $v \in \mathbb{S}$ we have

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

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

---

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

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

Corollary 3.30 (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*

$$\mathscr{A}^* \mathscr{A} \tilde{x} = \mathscr{A}^* y. \tag{3.7}$$

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 3.29 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 3.13,

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

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 3.29 leads to the uniqueness of $\tilde{x}$ 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 square solution in (3.7) is therefore unique. ∎

The beauty here is that (3.7) 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 3.47). 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 3.29.

Example 3.31 (linear least squares in finite dimensions). Consider the operator $\mathscr{A}$ defined in Example 3.18 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 3.30, 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 3.18. Note that we can reformulate the operator form $\|\mathscr{A}x - \boldsymbol{y}\|_{\mathbb{R}^2}$ by an equivalent matrix representation form. Indeed, let $\mathcal{A}$ 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 $\|\mathscr{A}x - \boldsymbol{y}\|_{\mathbb{R}^2} = \|\mathcal{A}\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^2}$. Again, $\mathcal{A}$ is not injective and a solution is not unique.

Example 3.32 (linear least squares with matrices). Consider $\mathcal{A} : \mathbb{F}^n \to \mathbb{F}^m$ and $\boldsymbol{y} \in \mathbb{F}^m$. Applying Corollary 3.30 we have that there exists $\tilde{\boldsymbol{x}} \in \mathbb{F}^n$ minimizing $\|\mathcal{A}\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{F}^m}$ and

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

When $\mathcal{A}$ is injective or equivalently full column rank, we have the uniqueness of $\tilde{\boldsymbol{x}}$ again by Corollary 3.30. Another way to see this is that in this case $\mathcal{A}^*\mathcal{A}$ is an invertible matrix which implies the uniqueness of the minimizer $\tilde{\boldsymbol{x}}$.

3.4. Application to the singular value decomposition (SVD). 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 3.25 states that $\exists \lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_n$ 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,

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

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

$$(3.8) \qquad\qquad\qquad\qquad \mathscr{A}^*\mathscr{A}u_i = \sigma_i^2 u_i.$$

Theorem 3.33 (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 $\sigma_1 \geqslant \ldots \geqslant \sigma_i \geqslant \cdots \geqslant \sigma_k \geqslant 0$ 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*

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

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

Proof. Starting from (3.8), let r be the maximum index for which $\sigma_r > 0$ and define $v_i = \frac{1}{\sigma_i}\mathscr{A}u_i$ for $i \leqslant r$, so that

$$(3.9) \qquad\qquad\qquad\qquad \mathscr{A}u_i = \sigma_i v_i.$$

Substituting (3.9) into (3.8) gives

$$(3.10) \qquad\qquad\qquad\qquad \mathscr{A}^* v_i = \sigma_i u_i.$$

We claim that $\{\sigma_i^2, v_i\}$ for $i \leqslant r$ are the eigenpairs of the self-adjoint operator $\mathscr{A}\mathscr{A}^*$. To see this, applying $\mathscr{A}$ to both sides of (3.10) to arrive at

$$(3.11) \qquad\qquad\qquad\qquad \mathscr{A}\mathscr{A}^* v_i = \sigma_i \mathscr{A}u_i = \sigma_i^2 v_i.$$

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 3.25. The third assertion is clear owing to the first assertion and (3.5). ∎

Example 3.34 (SVD for the closed range theorem, the rank-nullity theorem, and the fundamental theorem of linear algebra). The SVD decomposition in Theorem 3.33 allows us to provide trivial proofs of various important results in finite dimensions including the closed range Theorem 3.13, 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

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

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 yield

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

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

Similarly, we have

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

and

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

With these conclusions in hand, the assertions in the closed range Theorem 3.13 and Corollary 3.14 are now trivial.

The SVD also provides an obvious proof for the rank-nullity theorem [7] as

$$(3.12a) \qquad \underbrace{\dim(\mathbb{X})}_{=n} = \underbrace{\dim(\mathsf{N}(\mathscr{A}))}_{=n-r} + \underbrace{\dim(\mathsf{R}(\mathscr{A}))}_{=r},$$

and

$$(3.12b) \qquad \underbrace{\dim(\mathbb{Y})}_{=m} = \underbrace{\dim(\mathsf{N}(\mathscr{A}^*))}_{=m-r} + \underbrace{\dim(\mathsf{R}(\mathscr{A}^*))}_{=r},$$

which, together with the closed range Theorem 3.13, is the basis for the fundamental theorem of linear algebra [73]. This is demonstrated in Figure 1 which shows the important role of the four fundamental subspaces $\mathsf{R}\left(\mathscr{A}\right), \mathsf{N}\left(\mathscr{A}\right), \mathsf{R}\left(\mathscr{A}^*\right)$, and $\mathsf{N}\left(\mathscr{A}^*\right)$ 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 emphasize that if we remove the dimensions and assume that $\mathsf{R}\left(\mathscr{A}\right)$ is closed, then the two diagrams in Figure 1 also hold for infinite dimensional Hilbert spaces.

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

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

Figure 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.

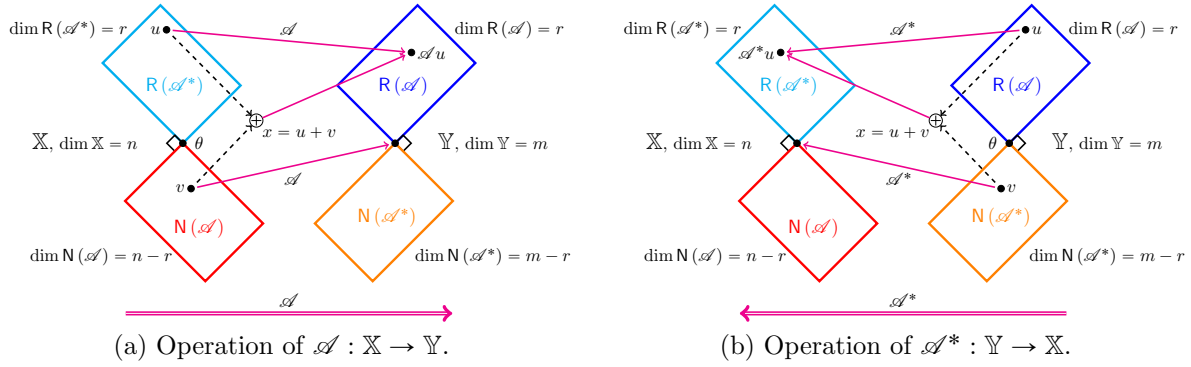Clearly, at the heart of the SVD is the eigenvalue decomposition (3.11), which could be challenging if it is analytically not tractable on the original operators. In that case, one has to resort to numerical methods. For finite-dimensional settings, an easier path is to explore the matrix representation of the linear operator.

Corollary 3.35 (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 $1 \leqslant i \leqslant k = min\{n, \ m\}$ 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 3.33. Denote $\boldsymbol{u}$ and $\boldsymbol{v}$ as the coordinate vectors of u and v in the bases $\boldsymbol{E}$ and $\boldsymbol{G}$, respectively, and $\mathcal{A}$ 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 $\mathcal{A}$ with*

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

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

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

Conversely, if $\{\sigma_i, \boldsymbol{u}_i, \boldsymbol{v}_i\}$, with $1 \leqslant i \leqslant k = min\{n, \ m\}$, are the singular triplets of $\mathcal{A}$, 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, Proposition 3.9, 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 3.36 (SVD of matrices). For a matrix $\mathcal{A} : \mathbb{X} = \mathbb{R}^n \to \mathbb{Y} = \mathbb{R}^m$ 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 $\mathcal{A}$ is itself and thus the SVD of $\mathcal{A}$ is given by Corollary 3.35. In this case, $\mathcal{A}^* = \mathcal{A}^T$. Furthermore: i) $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_n\}$ are orthonormal eigenvectors of $\mathcal{A}^T\mathcal{A}$; ii) $\{\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_m\}$ are orthonormal eigenvectors of $\mathcal{A}\mathcal{A}^T$; iii) $\sigma_i^2, \quad i = 1, \ldots r$, are nonzero eigenvalues of $\mathcal{A}^T\mathcal{A}$ or $\mathcal{A}\mathcal{A}^T$; and iv) from $\mathcal{A} = \sum_{i=1}^r \sigma_i \boldsymbol{v}_i \boldsymbol{u}_i^*$ we can write the full SVD form

$$\mathcal{A} = \underbrace{\left[\begin{array}{cccccc} | & | & & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_r & \ldots & \boldsymbol{v}_m \\ | & | & & | & & | \end{array} \right]}_{\mathcal{V}} \underbrace{\left[\begin{array}{ccccc} \sigma_1 & & & & \\ & \sigma_2 & & & \\ & & \ddots & & \\ & & & \sigma_r & \\ & & & & 0 \\ & & & & & \ddots \end{array} \right]}_{\Sigma} \underbrace{\left[\begin{array}{cccccc} | & | & & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_r & \ldots & \boldsymbol{u}_n \\ | & | & & | & & | \end{array} \right]^T}_{\mathcal{U}^T},$$

that is,

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

or the reduced SVD form

$$\mathcal{A} = \underbrace{\left[\begin{array}{cccc} | & | & & | \\ \boldsymbol{v}_1 & \boldsymbol{v}_2 & \ldots & \boldsymbol{v}_r \\ | & | & & | \end{array} \right]}_{\mathcal{V}_r} \underbrace{\left[\begin{array}{cccc} \sigma_1 & & & \\ & \sigma_2 & & \\ & & \ddots & \\ & & & \sigma_r \end{array} \right]}_{\Sigma_r} \underbrace{\left[\begin{array}{cccc} | & | & & | \\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & \ldots & \boldsymbol{u}_r \\ | & | & & | \end{array} \right]^T}_{\mathcal{U}_r^T},$$

that is,

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

Example 3.37. 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 3.6. We are going to find the singular value decomposition of $\mathscr{A}$ indirectly via its matrix representation using Corollary 3.35. Clearly, two orthonormal bases for $\mathbb{U}$ and $\mathbb{R}^2$ are $\boldsymbol{E} = \{1, x, \frac{1}{2}(3x^2 - 1)\}$, $\boldsymbol{G} = \{[1, 0]^T, [0, 1]^T\}$, respectively. It is a simple exercise to show that the matrix representation $\mathcal{A}$ of $\mathscr{A}$ in these two bases is given by

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

and from Proposition 3.9 we know that the matrix presentation $\mathcal{A}^*$ of the adjoint $\mathscr{A}^*$ is $\mathcal{A}^* = \mathcal{A}^T$. From the proof of Theorem 3.33, by computing the eigendecomposition of $\mathcal{A}\mathcal{A}^T$ and $\mathcal{A}^T\mathcal{A}$, we can find the full SVD of $\mathcal{A}$ as

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

Now, Corollary 3.35 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.

3.5. Application of adjoint to optimization with equality constraints.
The field of optimization is vast (see, e.g., [58, 10, 62, 9] and the references therein) and we restrict ourselves to unconstrained optimization problems and constrained optimization problems with equality constraints. We begin our development with unconstrained optimization in finite-dimensional spaces. Let $f : \mathbb{R}^n \ni \boldsymbol{x} \mapsto f(\boldsymbol{x}) \in \mathbb{R}$ and we are interested in studying the optimization problem $\min_{\boldsymbol{x} \in \mathbb{R}^N} f(\boldsymbol{x})$. 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 3.38. *y is a (local) minimizer of $f(x)$ if there exists a open neighborhood, i.e. $B_\delta(y) := \{x : |x - y| < \delta\}$ (ball with radius δ in $\mathbb{R}$), for some $\delta > 0$, such that*

$$f(x) \geqslant f(y), \quad \forall x \in B_\delta(y).$$

We are interested in finding the necessary and sufficient conditions for y to be a minimizer. To that end, we consider the Taylor remainder theorem [5, 47] which states that for twice-differentiable function $f(x)$ and for any $\varepsilon \in \mathbb{R}$, there exists $0 < \theta < 1$ such that

$$(3.13) \qquad f(y + \varepsilon) = f(y) + \varepsilon \underbrace{f'(y)}_{\text{gradient } g(y)} + \frac{1}{2}\varepsilon^2 \underbrace{f''(y + \theta\varepsilon)}_{\text{Hessian } h(y + \theta\varepsilon)}.$$

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

Lemma 3.39 (First and second order necessary conditions for optimality in $\mathbb{R}$). *Suppose $f(x) : \mathbb{R} \to \mathbb{R}$ is twice continuously differentiable in a neighborhood of a minimizer y. It is necessary that*
i) the gradient vanishes, i.e., $g(y) = 0$, and
ii) the Hessian is non-negative, i.e., $h(y) \geqslant 0$.

Proof. We carry out the proof by contradiction. For the first assertion, we suppose that $f'(y) < 0$ and note that we can pick[8] $\varepsilon > 0$ such that

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

and together with (3.13) we conclude that $f(y + \varepsilon) < f(y)$: a contradiction. A similar contradiction argument can be carry out if $f'(y) > 0$. Thus $f'(y) = 0$.

---

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

For the second assertion, suppose $h(y) = f''(y) < 0$. By continuity of $f''(x)$ we can choose sufficiently small $|\varepsilon|$ such that $f''(y + \theta\varepsilon) < 0$. Then (3.13) reduces to

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

which is a contradiction, and this concludes the proof. ∎

Corollary 3.40. *Suppose $f(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}$ is twice continuously differentiable in a neighborhood of a minimizer $\boldsymbol{y}$, where $\mathbb{R}^n$ is equipped with the standard Euclidean inner product. It is necessary that*

i) the gradient vanishes, i.e., $\frac{\partial f}{\partial \boldsymbol{x}_i}(\boldsymbol{y}) = 0$, $i = 1, \ldots, n$ and

ii) the Hessian matrix is semi-positive definite, i.e., $\mathcal{H}(\boldsymbol{y}) \geqslant 0$.

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 3.45 to derive the first order optimality condition in general vector spaces.

$$\boldsymbol{y} \text{ is a minimizer of } f(\boldsymbol{x})$$
$$\Downarrow \qquad \qquad \text{by definition}$$
$$f(\boldsymbol{x}) \geqslant f(\boldsymbol{y}), \quad \forall \boldsymbol{x} \in B_\delta(\boldsymbol{y}) := \{\boldsymbol{x} \in \mathbb{R}^n : \|\boldsymbol{x} - \boldsymbol{y}\|_{\mathbb{R}^n} < \delta\}$$
$$\Downarrow \qquad \qquad \text{pick an arbitrary } \boldsymbol{v}$$
$$F(\varepsilon) := f(\boldsymbol{y} + \varepsilon\boldsymbol{v}) \geqslant f(\boldsymbol{y}), \quad \forall \varepsilon \in B_{\delta/\|\boldsymbol{v}\|_{\mathbb{R}^n}}(0) := \{\varepsilon : |\varepsilon| < \delta/\|\boldsymbol{v}\|_{\mathbb{R}^n}\}$$
$$\Downarrow \qquad \qquad \text{by definition}$$
$$0 \text{ is a minimizer of } F(\varepsilon)$$
$$\Downarrow \qquad \qquad \text{by Lemma 3.39}$$
$$\frac{dF}{d\varepsilon}\big|_{\varepsilon=0} = 0$$
$$\Downarrow \qquad \qquad \text{by chain rule}$$
$$\sum_{i=1}^{n} \frac{\partial f}{\partial \boldsymbol{x}_i}(\boldsymbol{y}) \boldsymbol{v}_i = 0$$
$$\Downarrow \qquad \qquad \boldsymbol{v} \text{ is arbitrary}$$
$$\frac{\partial f}{\partial \boldsymbol{x}_i}(\boldsymbol{y}) = 0, \quad i = 1, \ldots, n.$$

∎

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 $\mathbb{X}^*$ consisting of linear and bounded functionals on $\mathbb{X}$. For $\ell \in \mathbb{X}^*$ and $u \in \mathbb{X}$, we use the standard duality pairing

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

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

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

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., [58, 6]).

Definition 3.41. *Suppose that there is a linear and bounded map* $\mathscr{D}f(u,\cdot) : \mathbb{X} \to \mathbb{R}$ *such that*

$$(3.14) \qquad f(u+v) = f(u) + \mathscr{D}f(u,v) + o(\|v\|_{\mathbb{X}}),$$

where the little-oh notation means

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

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 $\mathbb{X}^*$ and thus we interchangeably write it in the duality pairing form

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

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

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

Definition 3.42 (Fréchet gradient). *Let* $f : \mathbb{X} \to \mathbb{R}$. *The gradient of* $f(\cdot)$ *at* u, *denoted as* $\nabla f(u) \in \mathbb{X}$, *is defined as a function on* $\mathbb{X}$ *such that*

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

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

With Definition 3.42 at hand, we can identify the gradient of the Fréchet derivative and we will explore this fact in many results below.

Example 3.43. Consider $f : \mathbb{X} \to \mathbb{R}$ where $\mathbb{X} \equiv \mathbb{R}^n$ and $\mathbb{R}^n$ is endowed with a weighted inner product $(\boldsymbol{x}, \boldsymbol{y})_{\mathbb{R}^n, \mathcal{M}} = \boldsymbol{x}^T \mathcal{M} \boldsymbol{y}$ with $\mathcal{M}$ being a symmetric positive definite matrix. Suppose that the (classical) partial derivatives $\frac{\partial f}{\partial \boldsymbol{x}_i}$, $i = 1, \ldots, n$, of the f are continuous. From (3.14), it is easy to see that the Fréchet derivative can be written as

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

which, together with Definition 3.42, gives

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

We observe that the Fréchet derivative is a special case of the directional derivative, and when $\mathcal{M}$ is the identity matrix usual gradient vector $\left[\frac{\partial f}{\partial \boldsymbol{x}_1}, \ldots, \frac{\partial f}{\partial \boldsymbol{x}_n}\right]^T$ 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 $c : \mathbb{X} \ni u \mapsto c\left(u\right) \in \mathbb{Y}$, then the Fréchet derivative $\mathscr{D}c\left(u\right)$, when exists, can be computed as

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

The difference is now that $\mathscr{D}c\left(u\right)$ is a linear and bounded map from $\mathbb{X}$ to $\mathbb{Y}$, that is, $\mathscr{D}c\left(u\right) \in \mathfrak{B}\left(\mathbb{X}, \mathbb{Y}\right)$.

Example 3.44. Consider a vector-valued function $\mathbf{c}\left(\boldsymbol{x}\right) : \mathbb{R}^n \to \mathbb{R}^m$ where both $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the standard Euclidean inner products. Applying Example 3.43 for each component of $\mathbf{c}_i$, $i = 1, \ldots, m$ we have

$$\mathscr{D}\boldsymbol{c}\left(\boldsymbol{x}\right)\boldsymbol{v} = \begin{bmatrix} \frac{\partial \boldsymbol{c}_1}{\partial \boldsymbol{x}_1} & \frac{\partial \boldsymbol{c}_1}{\partial \boldsymbol{x}_2} & \cdots & \frac{\partial \boldsymbol{c}_1}{\partial \boldsymbol{x}_n} \\ \frac{\partial \boldsymbol{c}_2}{\partial \boldsymbol{x}_1} & \frac{\partial \boldsymbol{c}_2}{\partial \boldsymbol{x}_2} & \cdots & \frac{\partial \boldsymbol{c}_2}{\partial \boldsymbol{x}_n} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial \boldsymbol{c}_m}{\partial \boldsymbol{x}_1} & \frac{\partial \boldsymbol{c}_m}{\partial \boldsymbol{x}_2} & \cdots & \frac{\partial \boldsymbol{c}_m}{\partial \boldsymbol{x}_n} \end{bmatrix}\boldsymbol{v},$$

which, together with Definition 3.42, we can define

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

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

Lemma 3.45 (First order optimality condition for unconstrained optimization). *Suppose that $f : \mathbb{X} \to \mathbb{R}$ attains its extremum at u. Then it is necessary that*

$$(3.15) \qquad\qquad \mathscr{D}f\left(u\right)v = 0, \quad \forall v \in \mathbb{X},$$

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

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

by the Riesz representation Theorem 3.1.

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

$$f(v) \geqslant f(u), \quad \forall v \in B_\delta(u) := \{w \in \mathbb{X} : \|w - u\|_{\mathbb{X}} < \delta\},$$

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

(3.16) $$f(u + \varepsilon v) \geqslant f(u), \quad \forall \varepsilon \in B_{\delta/\|v\|_{\mathbb{X}}}(0).$$

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 3.38, inequality (3.16) is equivalent to saying that $F(\cdot)$ attains its minimum at $\varepsilon = 0$. Thus, from the first result of Lemma 3.39, we have

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

but this is equivalent to $\mathscr{D}f(u, v) = 0$ by Definition 3.41 of Fréchet derivative. ∎

Example 3.46 (First order optimality condition for unconstrained optimization in $\mathbb{R}^n$). Back to Example 3.43. Suppose that f attains its minimum at $\boldsymbol{x}$. Combining (3.15), the gradient found in Example 3.43, and Definition 3.42 yields

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

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

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

Example 3.47. We now revisit the least squares problem in Corollary 3.30 in the equivalent form: $\inf_{x \in \mathbb{X}} \frac{1}{2} (\mathscr{A}x - y, \mathscr{A}x - y)_{\mathbb{Y}}$. Using Definition 3.41, the first order optimality condition (3.15) reads

$$2 (\mathscr{A}v, \mathscr{A}x - y)_{\mathbb{Y}} = 0, \quad \forall v \in \mathbb{X},$$

that is,

$$\mathscr{A}^* \mathscr{A} x = \mathscr{A}^* y,$$

which is consistent with the least squares solution in (3.7).

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

$$\min_{u \in \mathbb{X}} f(u), \quad \text{subject to } c(u) = \theta, \text{ where } c(\cdot) : \mathbb{X} \to \mathbb{Y}.$$

If there were no constraint $c(u) = \theta$, then from Lemma 3.45 the optimality condition would be

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

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.,

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

Suppose c is Fréchet differentiable, it is therefore necessary that

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

To rigorously establish this result, we need the inverse function theorem [58], which in turn is a direct consequence of the implicit function theorem [49, 28].

Theorem 3.48 (Implicit function theorem). *Let $c : \mathbb{X} \times \mathbb{Z} \to \mathbb{Y}$ be continuously Fréchet differentiable and $\mathscr{D}_u c\left(u, z\right) : \mathbb{X} \to \mathbb{Y}$, is invertible at a point $\left[u_0, z_0\right]^T$, at which $c\left(u_0, z_0\right) = 0$. Then, there exist a neighborhood $B_\delta\left(z_0\right)$ and a continuously Fréchet differential function $g : B_\delta\left(z_0\right) \to \mathbb{X}$ such that $c\left(g\left(z\right), z\right) = 0$ for all $z \in B_\delta\left(z_0\right)$.*

Theorem 3.49 (Inverse function theorem). *Let $f : \mathbb{X} \to \mathbb{Z}$. Assume that $\mathscr{D}f\left(u_0\right)$ is continuous and maps $\mathbb{X}$ **onto** $\mathbb{Z}$. Then, there is a neighborhood $B_\delta\left(f\left(u_0\right)\right)$ of $f\left(u_0\right)$ such that $f\left(u\right) = z$ has a unique continuously differentiable solution $u(z)$ for every $z \in B_\delta\left(f\left(u_0\right)\right)$.*

Proof. The result is clear if we define $c\left(u, z\right) = z - f\left(u\right)$ and set $z_0 = f\left(u_0\right)$. Then by the implicit function Theorem 3.48, there exists $g : B_\delta\left(z_0\right) \to \mathbb{X}$ such that $0 = c\left(g\left(z\right), z\right) = z - f\left(g\left(z\right)\right) = 0$ for all $z \in B_\delta\left(z_0\right)$. Setting $u = g\left(z\right)$ concludes the proof. ∎

Lemma 3.50 (First order optimality condition for equality constraints). *Suppose $f : \mathbb{X} \to \mathbb{R}$ attains its extremum at u_0 subject to the constraint $c\left(u\right) = 0$, where $c : \mathbb{X} \to \mathbb{Y}$. Assume that both f and c are continuously Fréchet differentiable in an open set containing u_0, and $\mathscr{D}c\left(u_o\right)$ maps $\mathbb{X}$ **onto** $\mathbb{Y}$. Then, it is necessary that*

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

or equivalently

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

Proof. We follow closely the proof by contradiction in [58, Lemma 1 of Chapter 9]. Without lost of generality, asssume u_0 is a minimizer. Let us consider the transformation $g\left(u\right) = \left(f\left(u\right), c\left(u\right)\right) : \mathbb{X} \to \mathbb{R} \times \mathbb{Y}$. Assume that there exists h such that $\mathscr{D}c\left(u_0; h\right) = 0$ but $\mathscr{D}f\left(u_0; h\right) \neq 0$. Then the function $\left(\mathscr{D}f\left(u_0\right), \mathscr{D}c\left(u_0\right)\right)$ maps $\mathbb{X}$ onto $\mathbb{R} \times \mathbb{Y}$ since $\mathscr{D}c\left(u_0\right)$ maps $\mathbb{X}$ onto $\mathbb{Y}$. By the inverse function Theorem 3.49 there exists ε and u with $\|u - u_0\|_{\mathbb{X}} < \varepsilon$ such that $g\left(u\right) = \left(f\left(u_0\right) - \delta, 0\right)$ for some small $\delta > 0$. Thus, $f\left(u\right) = f\left(u_0\right) - \delta < f\left(u_0\right)$: a contradiction. ∎

Example 3.51. Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable and $\mathcal{A} \in \mathbb{R}^{m \times n}$, where $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the standard Euclidean inner products. We consider the following problem

$$\min_{\boldsymbol{x} \in \mathbb{R}^n} f\left(\boldsymbol{x}\right), \quad \text{subject to } \mathcal{A}\boldsymbol{x} = \boldsymbol{b}.$$

From Lemma 3.50, Example 3.43, and Example 3.44 we can write the the first order optimality condition as

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

or equivalently

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

i.e., *due to the constraint, the gradient of f at an optimum $\boldsymbol{x}$ does not vanish but is orthogonal to the nullspace of the gradient $\mathcal{A}$ 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 $\mathcal{Z}$ with columns comprising a basis of the nullspace of $\mathcal{A}$, then $\boldsymbol{v} = \mathcal{Z}\mathbf{r}$ 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

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

Note that $\mathbf{g}_r(\boldsymbol{x}) := \nabla f^T (\boldsymbol{x}) \, \mathcal{Z}$—the coordinates of the gradient in the nullspace of the constraint gradient—is known as the reduced gradient [32, 44]. 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 3.58. One of the reasons for its name is that its dimension is smaller than the dimension n of the original gradient vector $\nabla f(\boldsymbol{x})$. *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 3.40 we know that the derivative of the reduced gradient $\mathbf{g}_r$, namely the reduced Hessian $\mathcal{H}_r$, is necessary to be semi-positive definite in any direction in the reduced space. By the chain rule we have

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

from which it follows that

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

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

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

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

$$\mathcal{Z} = \left[\begin{array}{c} -\mathcal{U}^{-1}\mathcal{V} \\ I \end{array} \right],$$

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

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 3.13, the Lagrangian multiplier theorem [58, 10, 62, 9] is a straightforward equivalence to Lemma 3.50.

Theorem 3.52 (Lagrangian multiplier theorem). *Assume that $f : \mathbb{X} \to \mathbb{R}$ is continuously Fréchet differentiable and it attains the extremum at u_0 subject to the constraint $c(u) = 0$, where $c : \mathbb{X} \to \mathbb{Y}$ is continuously Fréchet differentiable. Suppose that $\mathscr{D}c(u_0)$ maps $\mathbb{X}$ **onto** $\mathbb{Y}$. Then, there exists an element $y \in \mathbb{Y}$ such that the following Lagrangian functional*

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

is stationary at u_0, i.e.,

$$(3.17) \qquad \mathscr{D}L(u_0)h = \mathscr{D}f(u_0)h + (y, \mathscr{D}c(u_0)h)_{\mathbb{Y}} = 0, \quad \forall h \in \mathbb{X},$$

or equivalently

$$(3.18) \qquad \nabla L(u_0) = \nabla f(u_0) + [\mathscr{D}c(u_0)]^* y = 0.$$

Proof. From Lemma 3.50 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 $\mathbb{X}$ onto $\mathbb{Y}$, the range of $\mathscr{D}c(u_0)$ closed and the closed range Theorem 3.13 gives

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

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

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

and this ends the proof. ∎

Remark 3.53. The appealing feature of the Lagrangian approach in Theorem 3.52 is that the first order optimality condition (3.18) is the standard optimality condition for unconstrained problem in Lemma 3.45, 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 y, 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* (3.18). 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 subsection 3.6.

Example 3.54. Back to Example 3.51. Applying the Lagrangian multiplier Theorem 3.52 together with the Riesz representation Theorem 3.1, the equivalent first order optimality condition (3.18) reduces to

$$\nabla f(\boldsymbol{x}) + \mathcal{A}^T \boldsymbol{y} = 0,$$

which says that *at an optimum of a constrained optimization problem, the gradient of the objective function $f(\boldsymbol{x})$ 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 $f(\boldsymbol{x})$ at an optimum is orthogonal to the nullspace of the gradient of the constraints.

Example 3.55. Let $f : \mathbb{R}^n \to \mathbb{R}$ and $\mathbf{c}(\boldsymbol{x}) : \mathbb{R}^n \to \mathbb{R}^m$, where $\mathbb{R}^n$ and $\mathbb{R}^m$ are endowed with the standard Euclidean inner products. We consider the following optimization problem

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

From Lemma 3.50, Example 3.43, Example 3.44, and the Lagrangian multiplier Theorem 3.52 we can write the first order optimality condition either as

$$(3.19) \qquad\qquad \nabla f(\boldsymbol{x}) + \nabla \mathbf{c}^T(\boldsymbol{x}) \boldsymbol{y} = \boldsymbol{0},$$

or as

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

Note that unlike the linear problem, $\boldsymbol{v}$ is a function of $\boldsymbol{x}$ since the nullspace of $\nabla \mathbf{c}(\boldsymbol{x})$ depends on $\boldsymbol{x}$. As a result, *the reduced Hessian is different from that of the linear constraint counterpart in Example* 3.51 *as we now show.* To that end, we compute the Fréchet derivative of $\mathbf{g}$ in any direction $\mathbf{p}$ in the reduced space $\mathsf{N}(\nabla \mathbf{c}(\boldsymbol{x}))$:

$$(3.20) \qquad\qquad \mathbf{p}^T \mathcal{H} \mathbf{h} = \mathbf{p}^T \nabla f^2 \boldsymbol{v} + \nabla f^T \mathscr{D} \boldsymbol{v}(\boldsymbol{x}, \mathbf{p}).$$

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

$$(3.21) \qquad\qquad \mathbf{p}^T \nabla^2 c_i \boldsymbol{v} + \nabla c_i^T \mathscr{D} \boldsymbol{v}(\boldsymbol{x}, \mathbf{p}) = 0.$$

Combining (3.19)–(3.21) gives

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

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

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

As can be seen, the Hessians of the constraints (which is zero for linear constraint case in Example 3.51) contribute to the reduced Hessian. Unlike Example 3.51 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 $\mathbf{c}(\boldsymbol{x}) = \boldsymbol{b}$.

To the end of 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 subsection 3.6, training deep neural network with backpropagation is a special case of this class. PDE-constrained optimization problem is another special case as shown in subsection 4.4.

Corollary 3.56 (Optimization with special equality constraint). *Consider optimization problems that can be expressed in the following form*

$$\min_{u \in \mathbb{X}, z \in \mathbb{Z}} f(u, z), \quad \textit{subject to } c(u, z) = 0, \textit{ where } c(\cdot, \cdot) : \mathbb{X} \times \mathbb{Z} \to \mathbb{Y},$$

where the Fréchet derivative of the constraint with respect to u, i.e. $\mathscr{D}_u c(u, z) : \mathbb{X} \to \mathbb{Y}$, is invertible at an optimum $[u_0, z_0]^T$. The first order optimality condition (3.18), together with the constraint, can be written as

(3.22a)	$c(u_0, z_0) = 0,$	*Forward equation,*
(3.22b)	$\nabla_u f(u_0, z_0) + [\mathscr{D}_u c(u_0, z_0)]^* y = 0,$	*Adjoint equation,*
(3.22c)	$\nabla_z f(u_0, z_0) + [\mathscr{D}_z c(u_0, z_0)]^* y = 0,$	*Control equation,*

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

Proof. The proof is a straightforward application of (3.18) 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 (3.22a) is simply the derivative of the Lagrangian with respect to the adjoint variable y. The LHS of the adjoint equation (3.22b) is nothing more than the derivative of the Lagrangian with respect to u, and the LHS of the control equation (3.22c) 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 y 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 3.48 there exists $B_\delta(z_0)$ and a continuously differentiable function $g : B_\delta(z_0) \to \mathbb{X}$ such that $u = g(z)$ solve the constraint $c(u, z) = 0$ for any $z \in B_\delta(z_0)$. The objective function becomes $f(g(z), z)$, and thus a function of only z, for all $z \in B_\delta(z_0)$. The optimization variable is reduced to only z and this is known as the *reduced space approach* [32, 44]. 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 3.57 (Reduced gradient in constrained optimization with equality constraints). *With the same setting as in Corollary 3.56, there exists a neighborhood $B_\delta(z_0)$ such that the reduced gradient at any $z \in B_\delta(z_0)$ is given by*

$$(3.23) \qquad \nabla f(g(z), z) = \nabla_z f(u, z) + [\mathscr{D}_z c(u, z)]^* y,$$

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

(3.24a) $$c(u, z) = 0, \qquad \textit{Forward equation,}$$
(3.24b) $$\nabla_u f(u, z) + [\mathscr{D}_u c(u, z)]^* y = 0, \qquad \textit{Adjoint equation.}$$

Proof. Note that the invertibility of a linear operator $\mathscr{A}$ implies the invertibility[9] 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

$$(\nabla f(g(z), z), h)_{\mathbb{Z}}$$
$$\|$$
$$\langle \mathscr{D} f(g(z), z), h \rangle_{\mathbb{Z}}$$
$$\|$$
$$\langle \mathscr{D}_z f(u, z), h \rangle_{\mathbb{Z}} + \langle \mathscr{D}_u f(u, z), \mathscr{D}_z g(z) h \rangle_{\mathbb{X}}$$
$$\|$$
$$(\nabla_z f(u, z), h)_{\mathbb{Z}} + (\nabla_u f(u, z), \mathscr{D}_z g(z) h)_{\mathbb{X}}$$
$$\|$$
$$(\nabla_z f(u, z), h)_{\mathbb{Z}} - \left(\nabla_u f(u, z), [\mathscr{D}_u c(u, z)]^{-1} \mathscr{D}_z c(u, z) h \right)_{\mathbb{X}}$$
$$\|$$
$$(\nabla_z f(u, z), h)_{\mathbb{Z}} - \left([\mathscr{D}_z c(u, z)]^* [\mathscr{D}_u c(u, z)]^{-*} \nabla_u f(u, z), h \right)_{\mathbb{Z}}$$
$$\|$$
$$(\nabla_z f(u, z), h)_{\mathbb{Z}} + \left([\mathscr{D}_z c(u, z)]^* y, h \right)_{\mathbb{Z}}$$

Gradient Definition 3.42

Chain rule

Gradient Definition 3.42

Derivative of the constraint

Adjoint Definition 3.2

$[\mathscr{D}_u c(u, z)]^* y := -\nabla_u f(u, z)$

where, as in the first equality, the derivative of the constraint in the third equality is given by the chain rule[10]:

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

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

Remark 3.58. Note that in practice, approximating the minimum u_0 is typically done using a gradient descent algorithm and Lemma 3.57 shows that the reduced gradient can be computed in each iteration for a given z via three steps: 1) solve the *forward equation* (3.24a) for $u(z)$, 2) solve the *adjoint equation* (3.24b) for $y(u(z), z)$, and 3) substitute $u(z)$ and $y(u(z), z)$ into (3.23) to obtain the reduced gradient. Moreover, the adjoint equation is always linear in the adjoint variable y regardless the linear or nonlinear nature of the forward equation. Note that full space iteration based on the first order optimality condition (3.22) is also possible and can be consulted from [62] and the references therein.

---

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

[10]The chain rule for Fréchet deriviation can be derived from (3.14). Let $f : \mathbb{X} \ni u \mapsto f(u) \in \mathbb{Y}$ and $g : \mathbb{Z} \ni z \mapsto g(z) \in \mathbb{X}$. We have $f(g(z + \varepsilon h)) = f(g(z) + \varepsilon \mathscr{D} g(z) h + o(\varepsilon)) = f(g(z)) + \varepsilon \mathscr{D}_u f(g(z)) (\mathscr{D} g(z) h + \varepsilon^{-1} o(\varepsilon)) + o(\varepsilon)$. Thus, $\mathscr{D}_z f(g(z)) h = \lim_{\varepsilon \to 0} \frac{f(g(z+\varepsilon h)) - f(g(z))}{\varepsilon} = \mathscr{D}_u f(g(z)) \mathscr{D} g(z) h = \langle \mathscr{D}_u f(g(z)), \mathscr{D} g(z) h \rangle_{\mathbb{X}}$

Example 3.59. To appreciate the adjoint approach, let us apply Lemma 3.57 to identify the forward equation, the adjoint equation, and the reduced gradient of the following finite dimensional optimization problem

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

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 $\det\left(\nabla_{\boldsymbol{u}}\mathbf{c}\right)\neq 0, \forall \boldsymbol{u},\mathbf{z}$ so that the implicit function theorem allows us to compute $\boldsymbol{u}$ as a function of $\mathbf{z}$ from the constraint. Applying Lemma 3.57 the reduced gradient reads

$$(3.25) \qquad \nabla f = \nabla_{\mathbf{z}} f\left(\boldsymbol{u}_0,\mathbf{z}_0\right) + \left[\nabla_{\mathbf{z}}\mathbf{c}\left(\boldsymbol{u}_0,\mathbf{z}_0\right)\right]^T \boldsymbol{y},$$

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

$$(3.26a) \qquad\qquad\qquad \mathbf{c}\left(\boldsymbol{u}_0,\mathbf{z}_0\right)=\mathbf{0}, \qquad\qquad \text{Forward equation,}$$

$$(3.26b) \qquad \nabla_{\boldsymbol{u}} f\left(\boldsymbol{u}_0,\mathbf{z}_0\right) + \left[\nabla_{\boldsymbol{u}}\mathbf{c}\left(\boldsymbol{u}_0,\mathbf{z}_0\right)\right]^T \boldsymbol{y}=\mathbf{0}, \qquad\qquad \text{Adjoint equation.}$$

3.6. Application of adjoint to backpropagation in deep learning. In this section, we consider standard fully connected deep neural network and use the adjoint method in Example 3.59 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 [50, 69], and the history of backpropagation can be traced back to [54, 55]. The gradient is needed for gradient-based methods (see, e.g., [62] and the references therein) such as stochastic gradient descent [46, 66, 70]. The extension of the adjoint method for other type of neural networks such as ResNet [38] and CNN [35, 42] 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.*

Definition 3.60 (L-layer Neural network). *Given $n_\ell, s_0, s_1, \ldots, s_{n_\ell} \in \mathbb{N}$, an n_ℓ-layer neural network is defined as the following series of composition*

$$(3.27) \qquad \begin{aligned} &\textit{Input layer}: \boldsymbol{a}^0 - \boldsymbol{x} = \mathbf{0}, \\ &\textit{The ith layer}: \boldsymbol{a}^i - \sigma\left(\mathcal{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right) = \mathbf{0}, \quad i=1,\ldots,n_\ell, \end{aligned}$$

where $\boldsymbol{x}\in\mathbb{R}^{s_0}$; $\mathcal{W}^i\in\mathbb{R}^{s_i}\times\mathbb{R}^{s_{i-1}}$ and $\boldsymbol{b}^i\in\mathbb{R}^{s_i}$, $i=1,\ldots,n_\ell$, 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 $\boldsymbol{u} := \left[\boldsymbol{a}^0,\ldots,\boldsymbol{a}^{n_\ell}\right]^T$, $\mathbf{z} := \left[\mathcal{W}^1,\boldsymbol{b}^1,\ldots,\mathcal{W}^{n_\ell},\boldsymbol{b}^{n_\ell}\right]^T$, and $\boldsymbol{c}\left(\boldsymbol{u},\mathbf{z}\right)=\mathbf{0}$ as the concatenation of all the sub-equations in (3.27). For concreteness, let us consider the loss (objective) function to be:

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

where $\boldsymbol{a}^{obs}$ is a given data (label). The neural network training problem is exactly the constrained optimization problem in Example 3.59. Thus

- The forward equation (3.26a), by definition, are nothing more than the neural network description in (3.27). For example, the ith block of forward sub-equations (the ith layer equation) are

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

  the corresponding ith block of forward solution is $\boldsymbol{u}^i = \boldsymbol{a}^i$, and the ith block of parameter is $\mathbf{z}^i = \left[\mathbf{z}_{\mathcal{W}}^i, \mathbf{z}_{\boldsymbol{b}}^i\right]^T := \left[\mathcal{W}^i, \boldsymbol{b}^i\right]^T$. 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 3.59 hold.
- To unfold the adjoint equation (3.26b), we note that the whole adjoint vector $\boldsymbol{y}$ is the concatenation of the adjoint sub-vector $\boldsymbol{y}^i$ corresponding to the ith layer equation in (3.27), for $i = 0, \ldots, n_\ell$. Thus, the ith adjoint equation corresponds to the derivative with respect to $\boldsymbol{u}^i = \boldsymbol{a}^i$ in (3.26b), and it reads

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

  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 $\sigma'\left(\mathcal{W}^{i+1} \boldsymbol{a}^i + \boldsymbol{b}^{i+1}\right)$ is a vector in (3.28). Thus, (3.28) 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 $\boldsymbol{y}^i$, $i = 0, \ldots, n_\ell$.
- To unfold the control equation (3.25) 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 $\mathbf{z}^i = \left[\mathbf{z}_{\mathcal{W}}^i, \mathbf{z}_{\boldsymbol{b}}^i\right]^T$ in $\mathbf{z}$. For DNN, these derivatives are given as: for $i = 1, \ldots, n_\ell$,

$$
(3.29) \quad \begin{aligned}
\frac{\partial f}{\partial \mathcal{W}^i} &= -\left[\boldsymbol{y^i} \circ \sigma'\left(\mathcal{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right)\right] \left(\boldsymbol{a}^{i-1}\right)^T, \\
\frac{\partial f}{\partial \boldsymbol{b}^i} &= -\boldsymbol{y}^i \circ \sigma'\left(\mathcal{W}^i \boldsymbol{a}^{i-1} + \boldsymbol{b}^i\right).
\end{aligned}
$$

The **backpropagation** nature of the network gradient is now clearly seen from the gradient expressions in (3.29) and the adjoint equations (3.28). Indeed, in (3.29) we need the ith adjoint state $\boldsymbol{y}^i$ in order to compute the gradients with respect to the weights and biases in the ith layer. The ith adjoint state $\boldsymbol{y}^i$ in turn depends on the $(i + 1)$th adjoint state $\boldsymbol{y}^{i+1}$, which depends on the $(i + 2)$th adjoint state $\boldsymbol{y}^{i+2}$, etc, all the way to the last adjoint state $\boldsymbol{y}^{n_\ell}$ corresponding to the network output layer. In other words, using the adjoint equations (3.28) we backpropagate to compute the adjoint solution from the output layer to the ith layer, and then compute the gradients using (3.29). From the backpropagation point of view, $\boldsymbol{y}^i$, $i = 1, \ldots, n_\ell$ 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 3.58. 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.

3.7. Application of adjoint to the stability of ordinary differential equations. 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 [56], and we limit ourselves to autonomous systems of the form

$$(3.30) \qquad \dot{\boldsymbol{x}} := \frac{d\boldsymbol{x}}{dt} = \boldsymbol{f}\left(\boldsymbol{x}\right),$$

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 $\boldsymbol{0} \in \mathsf{G}$ is an equilibrium point, i.e. $\boldsymbol{f}\left(\boldsymbol{0}\right) = \boldsymbol{0}$. We use $\boldsymbol{x}_0$ to denote the initial condition and I to denote the maximal interval of existence for a solution of (3.30). All the norms $\|\cdot\|$ and inner products $(\cdot, \cdot)$ in this section are the standard Euclidean ones, and in this case Example 3.5 shows that the adjoint of a real matrix is simply its transpose. For matrices, the norm is the induced operator norm.

Definition 3.61 (Lyapunov stability). *The equilibrium point $\boldsymbol{0}$ 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 $\boldsymbol{x}\left(0\right) \leqslant \delta$, we have $\boldsymbol{x}\left(t\right) \leqslant \varepsilon$ for all $t \in I \cap \left(0, \infty\right)$.*

Theorem 3.62 (Lyapunov direct method). *If there exists an open neighborhood U of $\boldsymbol{0}$ and a continuous differentiable function V such that*
1. *$V\left(\boldsymbol{0}\right) = 0$ and $V\left(\boldsymbol{z}\right) > 0$ for all $\boldsymbol{z} \in \mathsf{U} \backslash \left\{\boldsymbol{0}\right\}$, and*
2. *$V_{\boldsymbol{f}}\left(\boldsymbol{z}\right) := \left(\nabla V\left(\boldsymbol{z}\right), \boldsymbol{f}\left(\boldsymbol{z}\right)\right) := \sum_{i=1}^{n} \frac{\partial V}{\partial \boldsymbol{z}_i} \boldsymbol{f}_i\left(\boldsymbol{z}\right) \leqslant 0$ for all $\boldsymbol{z} \in \mathsf{U}$.*
Then $\boldsymbol{0}$ is a stable equilibrium point of (3.30).

Proof. See [56, Theorem 5.2]. ∎

Definition 3.63 (Asymptotic stability). *The equilibrium $\boldsymbol{0}$ is attractive if there exists $\delta > 0$ such that for every $x_0 \in \mathsf{G}$ such that $\|\boldsymbol{x}_0\| \leqslant \delta$, then the solution $\boldsymbol{x}\left(t\right) \to \boldsymbol{0}$ as $t \to \infty$. We say $\boldsymbol{0}$ asymptotically stable (in the sense of Lyapunov) if it is both stable and attractive.*

Theorem 3.64 (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 $V\left(\boldsymbol{0}\right) = 0$ and $V\left(\boldsymbol{z}\right) > 0$ for all $\boldsymbol{z} \in \mathsf{U} \backslash \left\{\boldsymbol{0}\right\}$, and $V_{\boldsymbol{f}}\left(\boldsymbol{z}\right) \leqslant 0$ for all $\boldsymbol{z} \in \mathsf{U}$, and
ii $\boldsymbol{0}$ is the inverse image of $V_{\boldsymbol{f}}\left(\boldsymbol{z}\right) = 0$, i.e., $V_{\boldsymbol{f}}^{-1}\left(0\right) = \boldsymbol{0}$.
Then $\boldsymbol{0}$ is asymptotically stable.

Proof. See [56, 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 $\mathsf{G} = \mathbb{R}^n$, $\mathcal{A} \in \mathbb{R}^{n \times n}$, and

$$(3.31) \qquad\qquad \dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) = \mathcal{A}\boldsymbol{x}.$$

Clearly, the solution of the (3.31) can be written as matrix exponential [37, 61, 41]

$$\boldsymbol{x}(t) = \exp(\mathcal{A}t)\,\boldsymbol{x}_0.$$

Definition 3.65 (Exponential stability). *The equilibrium* **0** *is called exponentially stable if there exist $M \geqslant 1$ and $\alpha > 0$ such as*

$$\|\exp(\mathcal{A}t)\,\boldsymbol{x}_0\| \leqslant M \exp(-\alpha t)\,\|\boldsymbol{x}_0\|, \quad \forall t \geqslant 0 \text{ and } \forall \boldsymbol{x}_0 \in \mathbb{R}^n.$$

Definition 3.66 (Hurwitz matrices). *Let $\sigma(\mathcal{A})$ denote the spectrum (the collection of all eigenvalues) of $\mathcal{A}$. $\mathcal{A}$ is Hurwitz if $\sigma(\mathcal{A}) \subset \{\lambda \in \mathbb{C} : \Re(\lambda) < 0\}$.*

Proposition 3.67. *Let $\mathcal{A} \in \mathbb{R}^{n \times n}$. The following statements are equivalent:*

 i) *$\mathcal{A}$ is Hurwitz.*
 ii) *$\mathbf{0}$ is an exponentially stable equilibrium of (3.31).*
iii) *$\mathbf{0}$ is an asymptotically stable equilibrium of (3.31).*

Proof. See [56, Proposition 5.25]. ∎

We are in the position to discuss one of the main results of this section.

Theorem 3.68 (Necessary and sufficient conditions for exponential stability). *$\mathcal{A} \in \mathbb{R}^{n \times n}$ is Hurwitz iff for each symmetric positive definite (SPD) matrix $\mathcal{Q} \in \mathbb{R}^{n \times n}$, the matrix equation*

$$\mathcal{P}\mathcal{A} + \mathcal{A}^*\mathcal{P} + \mathcal{Q} = 0$$

has a SPD solution $\mathcal{P} \in \mathbb{R}^{n \times n}$.

Proof. For the necessity, suppose $\mathcal{A}$ is Hurwitz. It follows from Proposition 3.67 that **0** is an exponentially stable equilibrium, that is, there exists $M > 0$ and $\alpha > 0$ such that

$$\|\exp(\mathcal{A}t)\| = \sup_{\boldsymbol{x}_0 \in \mathbb{R}^n} \frac{\|\exp(\mathcal{A}t)\,\boldsymbol{x}_0\|}{\|\boldsymbol{x}_0\|} \leqslant M \exp(-\alpha t), \quad t \geqslant 0.$$

Now for any SPD matrix $\mathcal{Q}$, let us define

$$\mathcal{P} := \int_0^\infty \exp(\mathcal{A}^*)\,\mathcal{Q}\exp(\mathcal{A}t)\,dt,$$

which is a well-defined matrix since

$$\|\mathcal{P}\| \leqslant \int_0^\infty \|\exp(\mathcal{A}^*t)\|\,\|\mathcal{Q}\|\,\|\exp(\mathcal{A}t)\|\,dt \leqslant M\,\|\mathcal{Q}\|\int_0^\infty \exp(-2\alpha t)\,dt < \infty,$$

where in the second inequality we have used the fact from Proposition 3.3 that the norm of a linear continuous operator is equal to the norm of its adjoint. $\mathcal{P}$ is SPD as

$$(\boldsymbol{x}, \mathcal{P}\boldsymbol{x}) = \int_0^\infty (\boldsymbol{x}, \exp(\mathcal{A}^*t)\,\mathcal{Q}\exp(\mathcal{A}t)\,\boldsymbol{x})\,dt = \int_0^\infty (\exp(\mathcal{A}t)\,\boldsymbol{x}, \mathcal{Q}\exp(\mathcal{A}t)\,\boldsymbol{x})\,dt,$$

and the fact that $\mathcal{Q}$ is SPD. Furthermore,

$$\mathcal{P}\mathcal{A} + \mathcal{A}^*\mathcal{P} = \int_0^\infty \frac{d\left(\exp\left(\mathcal{A}^*t\right)\mathcal{Q}\exp\left(\mathcal{A}t\right)\right)}{dt}\, dt = -\mathcal{Q},$$

where we have used the fact that $\exp\left(\mathcal{A}t\right)$, and hence $\exp\left(\mathcal{A}^*t\right)$, decays exponentially to 0.

For the sufficiency, Proposition 3.67 says that we only need to show that $\mathbf{0}$ is an asymptotically stable equilibrium. To that end, let us construct the following function

$$V\left(\mathbf{z}\right) := \left(\mathbf{z}, \mathcal{P}\mathbf{z}\right).$$

Clearly $V\left(\mathbf{z}\right) \geqslant 0$ for all $\mathbf{z} \in \mathbb{R}^n$, and $V\left(\mathbf{z}\right) = 0$ iff $\mathbf{z} = \mathbf{0}$ as $\mathcal{P}$ is SPD. Furthermore

$$V_{\mathbf{f}}\left(\mathbf{z}\right) = 2\left(\mathcal{P}\mathbf{z}, \mathcal{A}\mathbf{z}\right) = \left(\mathbf{z}, \left(\mathcal{P}\mathcal{A} + \mathcal{A}^*\mathcal{P}\right)\mathbf{z}\right) = -\left(\mathbf{z}, \mathcal{Q}\mathbf{z}\right) \leqslant 0.$$

Thus, by Theorem 3.64, $\mathbf{0}$ is asymptotically stable, and this ends the proof. ∎

Let us now use Theorem 3.68 to study the stability of the equilibrium $\mathbf{0}$ of the general nonlinear system (3.30).

Hypothesis 3.69 (Nonlinearly purturbed linear ODE systems). We assume that $\mathbf{f}\left(\mathbf{x}\right) = \mathcal{A}\mathbf{x} + \mathbf{h}\left(\mathbf{x}\right)$ where $\mathcal{A} \in \mathbb{R}^{n\times n}$ and $\mathbf{h} : \mathsf{G} \to \mathbb{R}^n$ is continuous with

$$(3.32) \qquad\qquad\qquad \lim_{\mathbf{z}\to\mathbf{0}} \frac{\|\mathbf{h}\left(\mathbf{z}\right)\|}{\|\mathbf{z}\|} = 0,$$

that is, $\mathbf{h}\left(\mathbf{z}\right) = o\left(\mathbf{z}\right)$. In other words, $\mathbf{h}$ approaches $\mathbf{0}$ faster than $\mathbf{z}$.

Theorem 3.70 (Linear stability implies nonlinear stability). *Assume Hypothesis 3.69 holds. If $\mathbf{0}$ is an asymptotically stable equilibrium of (3.31), it is also an asymptotically stable equilibrium of (3.30).*

Proof. Suppose $\mathbf{0}$ is an asymptotically stable equilibrium of (3.31). Using Theorem 3.68 we can pick $\mathcal{Q} = I$, and form $V\left(\mathbf{z}\right) := \left(\mathbf{z}, \mathcal{P}\mathbf{z}\right)$. It follows that $V\left(\mathbf{0}\right) = 0$ and $V\left(\mathbf{z}\right) > 0$ for all $\mathbf{z} \neq \mathbf{0}$ owing to the SPD property of $\mathcal{P}$. In order to use Theorem 3.64 to conclude the proof, we just need to show that $V_{\mathbf{f}}\left(\mathbf{z}\right) < 0$. We have

$$V_{\mathbf{f}}\left(\mathbf{z}\right) = 2\left(\mathcal{P}\mathbf{z}, \mathcal{A}\mathbf{z} + \mathbf{h}\left(\mathbf{z}\right)\right) = -\|\mathbf{z}\|^2 + 2\left(\mathcal{P}\mathbf{z}, \mathbf{h}\left(\mathbf{z}\right)\right) \leqslant -\|\mathbf{z}\|^2 + 2\|\mathcal{P}\|\,\|\mathbf{z}\|\,\|\mathbf{h}\left(\mathbf{z}\right)\|,$$

where we have used $\mathcal{P}\mathcal{A} + \mathcal{A}^*\mathcal{P} = -I$ in the second equality. Next, using (3.32) we can pick a sufficient small neighborhood $\mathsf{U} = B_\varepsilon\left(\mathbf{0}\right)$ such that $\|\mathbf{h}\left(\mathbf{z}\right)\| \leqslant \frac{\|\mathbf{z}\|}{4\|\mathcal{P}\|}$. Thus $V_{\mathbf{f}}\left(\mathbf{z}\right) < 0$ for all $\mathbf{z} \in \mathsf{U}$, and this ends the proof. ∎

Theorem 3.70 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 3.68 and Theorem 3.70 that adjoint plays the key role in establishing a necessary and sufficient condition for the stability of an equilibrium of an ODE system.*

Example 3.71 (Linearization of nonlinear ODE systems). We assume $\boldsymbol{f} : \mathsf{G} \to \mathbb{R}^n$ is differentiable at $\boldsymbol{0}$, and $\boldsymbol{0} \in \mathsf{G}$ is an equilibrium. Let us set

$$\mathcal{A} := \nabla \boldsymbol{f}\left(\boldsymbol{0}\right),$$

where the gradient is defined in Definition 3.42 and Example 3.44. The definition of Fréchet derivative in (3.14) implies (3.32) with $\boldsymbol{h}\left(\boldsymbol{x}\right) := \boldsymbol{f}\left(\boldsymbol{x}\right) - \mathcal{A}\boldsymbol{x}$. Then by Theorem 3.70, the asymptotic stability of $\boldsymbol{0}$ for the linearized system implies the asymptotic stability of $\boldsymbol{0}$ for the original nonlinear system.

Example 3.72 (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 [64, 39, 14, 59, 27, 51, 23, 15, 40, 76]. 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 [14, 59, 27, 51, 23, 15, 40] in which, for example in an Susceptible-Exposed-Infectious-Recovered-Susceptible (SEIRS) model, the components of $\boldsymbol{x}$ in (3.30) 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 2 is our effort[11] in sketching the association, via the next generation matrix approach [43, 76, 25], of the reproduction number being less than unity and the asymptotic stability of $\boldsymbol{0}$ as a DFE of an abstract epidemic ODE model $\dot{\boldsymbol{x}} = \boldsymbol{f}\left(\boldsymbol{x}\right)$. Note that the top arrow is the implication (due to Theorem 3.70) and the rest are equivalences. (See our work [3], 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.

4. Part II: Adjoint operators in Infinite dimensional Hilbert spaces. In this second part, we shall consider linear mappings between infinite dimensionsal Hilbert spaces. Of our
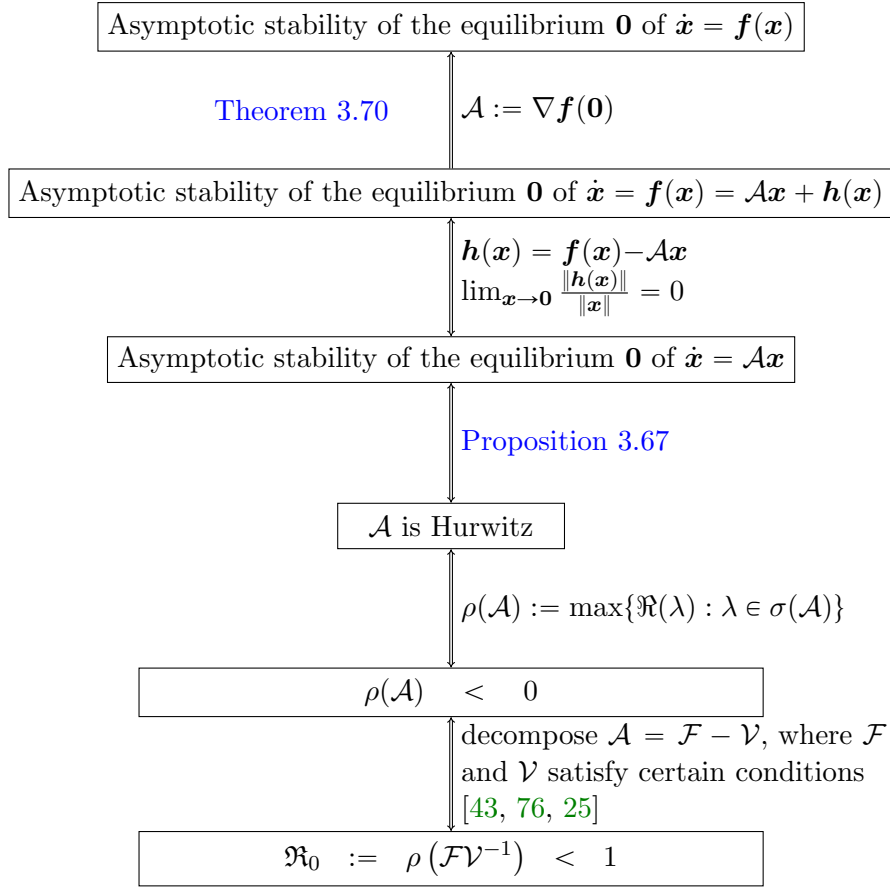
---

[11]Unpublished notes.

Asymptotic stability of the equilibrium $\mathbf{0}$ of $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x})$

Theorem 3.70 $\mathcal{A} := \nabla \boldsymbol{f}(\mathbf{0})$

Asymptotic stability of the equilibrium $\mathbf{0}$ of $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}) = \mathcal{A}\boldsymbol{x} + \boldsymbol{h}(\boldsymbol{x})$

$\boldsymbol{h}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) - \mathcal{A}\boldsymbol{x}$
$\lim_{\boldsymbol{x} \to \mathbf{0}} \frac{\|\boldsymbol{h}(\boldsymbol{x})\|}{\|\boldsymbol{x}\|} = 0$

Asymptotic stability of the equilibrium $\mathbf{0}$ of $\dot{\boldsymbol{x}} = \mathcal{A}\boldsymbol{x}$

Proposition 3.67

$\mathcal{A}$ is Hurwitz

$\rho(\mathcal{A}) := \max\{\Re(\lambda) : \lambda \in \sigma(\mathcal{A})\}$

$\rho(\mathcal{A}) \quad < \quad 0$

decompose $\mathcal{A} = \mathcal{F} - \mathcal{V}$, where $\mathcal{F}$
and $\mathcal{V}$ satisfy certain conditions
[43, 76, 25]

$\mathfrak{R}_0 \quad := \quad \rho\left(\mathcal{F}\mathcal{V}^{-1}\right) \quad < \quad 1$

Figure 2: 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 3.70) and the rest are equivalences.

particular interest are densely defined differential operators. A sufficient general definition of adjoint operator that could embrace a variety of problems is the following (see, e.g., [71]).

Definition 4.1 (Adjoint operators). *$\mathscr{A}^*$ is called the adjoint of $\mathscr{A} \in \mathfrak{L}(\mathbb{U}, \mathbb{V})$ iff*

(4.1) $$(\mathscr{A}u, v)_{\mathbb{V}} = (u, \mathscr{A}^*v)_{\mathbb{U}} \quad \forall u \in \mathsf{D}(\mathscr{A}) \ \text{and} \ \forall v \in \mathsf{D}(\mathscr{A}^*),$$

where

$$\mathsf{D}(\mathscr{A}^*) := \{v : \ \text{the map} \ u \mapsto (\mathscr{A}u, v)_{\mathbb{V}} \ \text{is continuous on} \ \mathbb{U}\}.$$

Clearly, Definition 4.1 reduces to Definition 3.2 when $\mathscr{A}$ is continuous and is defined on the whole space $\mathbb{U}$.

Example 4.2. 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

$$(4.2) \qquad (u,v)_{L^2(\Omega)} = \int_\Omega u(\boldsymbol{x})v(\boldsymbol{x})d\Omega.$$

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 $z \in \mathfrak{C}^1(\Omega) \subset \mathbb{L}^2(\Omega)$ is a distributed parameter over Ω with sufficiently smooth boundary $\partial\Omega$. We define the graph space $\mathbb{H}_\mathscr{A}$ as

$$\mathbb{H}_\mathscr{A} := \left\{ u \in \mathbb{H}^1(\Omega) : \mathscr{A}u \in \mathbb{L}^2(\Omega) \right\} = \left\{ u \in \mathbb{H}^1(\Omega) : -\nabla \cdot (e^z \nabla u) \in \mathbb{L}^2(\Omega) \right\}.$$

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 arugment and it is provided in Proof 41 of Appendix A.

Example 4.3. We now consider a weak formulation for the elliptic differential operator in Example 4.2. This will be important for studying the well-posedness of the associated partial differential equations in Example 4.31. 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

$$|a(u,v)| \leqslant \|e^z\|_{\mathbb{L}^\infty(\Omega)} \|\nabla u\|_{\mathbb{L}^2(\Omega)} \|\nabla v\|_{\mathbb{L}^2(\Omega)} \leqslant \|e^z\|_{\mathbb{L}^\infty(\Omega)} \|u\|_{\mathbb{H}^1(\Omega)} \|v\|_{\mathbb{H}^1(\Omega)},$$

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[12] $\mathscr{A} : \mathbb{H}_0^1(\Omega) \to \mathbb{H}_0^1(\Omega)$ as

$$(\mathscr{A}u, v)_{\mathbb{H}_0^1(\Omega)} := a(u,v).$$

---

[12]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 3.1 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 $a(u,v)$: $\|\mathscr{A}u\| = \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A}u, v)}{\|v\|_\mathbb{Y}} = \sup_{v \in \mathbb{Y}} \frac{a(u,v)}{\|v\|_\mathbb{Y}} \leqslant \beta \|u\|_\mathbb{X}$. 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 4.4. 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 (4.2). 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 $\boldsymbol{\beta} \in \left[\mathfrak{C}^1(\Omega, \mathbb{R})\right]^n$ 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 $\mathbb{H}_{\mathscr{A}} := \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) := \{u : u \in \mathbb{L}^2(\Omega) \text{ and } \boldsymbol{\beta} \cdot \nabla u \in \mathbb{L}^2(\Omega)\}$ which is dense in $\mathbb{L}^2(\Omega)$ under sufficient regularity[13] of the domain Ω [30]. The domain of $\mathscr{A}$ is defined as a subset of the graph space, namely, $\mathsf{D}(\mathscr{A}) := \{u \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) : \boldsymbol{\beta} \cdot \boldsymbol{n} u = 0 \text{ in } \partial\Omega_{\text{in}}\}$. 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 4.5 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

$$\mathsf{D}(\mathscr{A}^*) = \{v \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) : \boldsymbol{\beta} \cdot \boldsymbol{n} u = 0 \text{ in } \partial\Omega_{\text{out}}\}.$$

Example 4.5 (Friedrichs' systems). We are interested in Friedrichs' system that embraces a large class of elliptic, parabolic, hyperbolic, and mixed-type PDEs operators [33]:

(4.3) $$\mathscr{A}u := \sum_{k=1}^n \mathcal{A}_k \partial_k u + \mathcal{C}u \text{ in } \in \Omega,$$

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 $\mathcal{A}_k$ and $\mathcal{C}$ 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., [33, 30, 45, 31]):

(4.4a) $$\mathcal{C} \in [\mathbb{L}^\infty(\Omega)]^{m,m},$$

(4.4b) $$\mathcal{A}_k \in [\mathbb{L}^\infty(\Omega)]^{m,m}, \quad k = 1, \ldots, n, \quad \text{and} \quad \sum_{k=1}^n \partial_k \mathcal{A}_k \in [L^\infty(\Omega)]^{m,m},$$

(4.4c) $$\mathcal{A}_k = (\mathcal{A}_k)^T \text{ in } \Omega, \quad k = 1, \ldots, n,,$$

(4.4d) $$\mathcal{C} + \mathcal{C}^T + \sum_{k=1}^n \partial_k \mathcal{A}_k \geqslant 2\alpha_0 I \text{ in } \Omega,$$

---

[13]Assume Ω has segment property [4].

where $\alpha_0 > 0$ is some coercivity constant. In this paper, we consider the following abstract boundary condition for $\mathscr{A}$ [30]: let $\mathcal{D} := \sum_{k=1}^{n} \mathcal{A}_k \boldsymbol{n}_k$, where $\boldsymbol{n}_k$ is the kth component of the unit outward normal vector $\boldsymbol{n}$ on $\partial\Omega$, and assume there exists $\mathcal{M} \in [\mathbb{L}^\infty(\partial\Omega)]^{m,m}$ such that

$$(4.5a) \qquad\qquad\qquad\qquad \mathcal{M} + \mathcal{M}^T \geqslant 0 \text{ on } \partial\Omega,$$

$$(4.5b) \qquad\qquad\qquad\qquad (\mathcal{D} - \mathcal{M}) u = 0 \text{ on } \partial\Omega,$$

$$(4.5c) \qquad\qquad \mathsf{N}(\mathcal{D} - \mathcal{M}) + \mathsf{N}(\mathcal{D} + \mathcal{M}) = \mathbb{R}^m, \text{ on } \partial\Omega.$$

Following [29], we define[14] the graph space of $\mathscr{A}$ using its differential part $\mathscr{B} := \mathscr{A} - \mathcal{C}$:

$$\mathbb{H}_{\mathscr{A}} := \left\{ u \in \left[\mathbb{L}^2(\Omega)\right]^m : \mathscr{B}u \in \left[\mathbb{L}^2(\Omega)\right]^m \right\},$$

which is dense [31] in $\left[\mathbb{L}^2(\Omega)\right]^m$ when Ω is sufficiently regular (see Example 4.4 for an example). Furthermore, from the definition, it is easy to see that $\mathscr{A}$ is linear and continuous on $\mathsf{D}(\mathscr{A}) := \{x \in \mathbb{H}_{\mathscr{A}} : (\mathcal{D} - \mathcal{M}) x = 0 \text{ on } \partial\Omega\}$ 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 Proof 42 in Appendix A):

$$\mathscr{A}^* v = -\sum_{k=1}^{n} \partial_k (\mathcal{A}_k v) + \mathcal{C}^T v \in \mathbb{L}^2(\Omega),$$

for any $v \in \mathsf{D}(\mathscr{A}^*)$, where

$$\mathsf{D}(\mathscr{A}^*) = \left\{ v \in \mathbb{H}_{\mathscr{A}} : \mathcal{D}v \in [\mathsf{N}(\mathcal{D} - \mathcal{M})]^{\perp} \right\}.$$

Remark 4.6. Note that we have considered Friedrichs systems with full coercivity in Example 4.5. Applying such a general setting to various concrete PDEs [29, 30] will reveal concrete adjoints and their domains, but we omit the details here. The elliptic operator in Example 4.2, when written in the first order form, is a particular two-field Friedrichs system [29, 30], whose adjoint can be derived similarly.

4.1. Application of adjoint to ill-posed problems. In this section, we consider continuous linear operator defined on the whole space and we shall extend the spectral decomposition in Corollary 3.25 and SVD decomposition in Theorem 3.33 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.

Definition 4.7 (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.*

---

[14]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 [29, 30].

A direct consequence of Definition 4.7 is that any compact operator is a linear and continuous map.

Corollary 4.8. *If $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ is compact, then $\mathscr{A} \in \mathfrak{B}(\mathbb{X}, \mathbb{Y})$.*

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., [6, 16, 63, 67]). One of the important consequences is the Hilbert-Schmidt theorem (see, e.g., [6, 16, 63, 67, 72]), which is a generalization of Corollary 3.25.

Theorem 4.9 (Hilbert-Schmidt theorem for self-adjoint compact operators). *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*

$$(4.6) \qquad x = \sum_i (\varphi_i, x)_{\mathbb{X}} \, \varphi_i + \mathscr{P}x,$$

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 $\mathsf{R}(\mathscr{B})$.

Remark 4.10. The fact that $\mathscr{B}$ is compact when $\mathbb{X}$ is a finite dimensional space implies that Theorem 4.9 is a generalization of Corollary 3.25. 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 $\left\{\varphi_j^0\right\}_{j=1}^d$ and thus $\mathscr{P}x = \sum_{j=1}^d \left(\varphi_j^0, x\right)_{\mathbb{X}} \varphi_j^0$, which can be absorbed into the first sum on the right side of (4.6) 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 3.25.

We now follow the exposition in subsection 3.4 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 [63, 6]. The Hilbert-Schmidt Theorem 4.9 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 $\lambda_i \geqslant 0$. Let us define the *singular value* σ_i of $\mathscr{A}$ as

$$\sigma_i := \sqrt{\lambda_i}. \tag{4.7}$$

We are now in the position to study the singular value decomposition for compact operators (see, e.g.,[22]) that is a direct extension of Theorem 3.33.

Theorem 4.11 (Singular value decomposition for compact operators). *Let $\{\sigma_i\}$ be the sequence of non-zero singular values (defined in (4.7)) of a compact operator $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ and be ordered as*

$$\sigma_1 \geqslant \sigma_2 \geqslant \ldots,$$

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 3.33. The key that we exploit is the Hilbert-Schmidt theorem Theorem 4.9.
1. By Theorem 4.9 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 4.9 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,

$$\mathscr{A} x = \sum_i \mu_i \, (x, \varphi_i)_{\mathbb{X}} \, \phi_i.$$

$\blacksquare$

The next result [22], 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 4.12 (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) $\sum \frac{1}{\sigma_i^2} |(y, \phi_i)_{\mathbb{Y}}|^2 < \infty,$
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 4.11 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 3.13 [15] 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 4.9 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

$$\|x\|_{\mathbb{X}}^2 = \sum_i |(x, \varphi_i)_{\mathbb{X}}|^2 + \|\mathscr{P}x\|_{\mathbb{X}}^2,$$

---

[15]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 [68] 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 [63, 6]. 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.

which in turns implies

$$\sum_i |(x, \varphi_i)_{\mathbb{X}}|^2 \leqslant \|x\|_{\mathbb{X}}^2 < \infty.$$

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 4.9 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 4.8, 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 4.9 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. ∎

We now discuss the important consequence of the Picard Theorem 4.12, that is, *inverting a compact operator is an ill-posed problem*. We observe that

$$y = \mathscr{A} x = \sum_i \sigma_i (x, \phi_i)_{\mathbb{X}} \phi_i,$$

where we have used the second assertion of Theorem 4.11. 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

$$\tilde{x} = \sum_i \frac{1}{\mu_i} (\tilde{y}, \phi_i)_{\mathbb{Y}} \varphi_i = x + \frac{\delta}{\mu_N} \varphi_N.$$

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., [20, 19, 18] and the references therein).

Definition 4.13 (Well-posedness). *In Hadamard's sense [36], 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 4.14 (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

$$y(t) = \mathscr{A}x := \int_0^t x(s)\,ds, \quad 0 \leqslant t \leqslant 1,$$

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 $\mathbb{X} = \mathfrak{C}([0,1]), \mathbb{Y} = \mathfrak{C}([0,1])$. 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

$$\|\tilde{y} - y\|_{\mathfrak{C}([0,1])} := \sup_{t \in [0,1]} |\tilde{y}(t) - y(t)| \to 0, \quad \text{as } N \to \infty,$$

but

$$\|\tilde{x} - x\|_{\mathfrak{C}([0,1])} = \alpha \quad \forall N.$$

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[16] 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 $\mathfrak{C}([0,1])$ into $\mathfrak{C}^1([0,1]) \subset \mathfrak{C}([0,1])$. Since the inverse of a compact operator is unbounded, inverting the fundamental theorem of calculus is unstable by the Picard Theorem 4.12. The setting $\mathbb{X} = \mathfrak{C}([0,1]), \mathbb{Y} = \mathfrak{C}([0,1])$ thus leads to an ill-posed problem.

- Now let us consider $\mathbb{X} = \mathfrak{C}([0,1]), \mathbb{Y} = \mathfrak{C}^1([0,1])$. In this case we have

$$\|\tilde{y} - y\|_{\mathfrak{C}^1([0,1])} := \|\tilde{y} - y\|_{\mathfrak{C}([0,1])} + \|\tilde{y}' - y'\|_{\mathfrak{C}([0,1])} = \alpha, \quad \forall N,$$

and since

$$\|\tilde{x} - x\|_{\mathfrak{C}([0,1])} = \alpha \quad \forall N.$$

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

---

[16] By the Ascoli-Arzela theorem [63, 6].

surjectivity is also clear. Consequently, the inverse problem is well-posed in the Hadamard's sense.[17]

Remark 4.15. 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.

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 3.30:

$$(4.8) \qquad \min_x \frac{1}{2} \|\mathscr{A}x - y\|_{\mathbb{Y}}^2 .$$

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 (3.7) 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 (4.8) 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* [74, 75], which proposes to solve the following nearby problem

$$(4.9) \qquad \min_{x \in \mathbb{X}} \frac{1}{2} \|\mathscr{A}x - y\|_{\mathbb{Y}}^2 + \frac{\kappa}{2} \|x - x_0\|_{\mathbb{X}}^2 ,$$

where x_0 is some "prior" reference function and κ is known as the *regularization parameter*. To show that the *regularized optimization* problem (4.9) is well-posed, we need the projection theorem Theorem 3.29 and the following key result from the Riesz-Fredholm theory [21].

Lemma 4.16. *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 4.17. *For any $\kappa > 0$, the regularized optimization problem (4.9) is well-posed.*

Proof. Without loss of generality, assume $\kappa = 1$. We begin by rewrite the optimization (4.9) into the following equivalent form

$$(4.10) \qquad \min_z \frac{1}{2} \|\mathscr{B}z - w\|_{\mathbb{Y} \times \mathbb{X}}^2 ,$$

---

[17]Note that this is an instance of the Tikhonov theorem [26] since $\mathfrak{C}^1([0,1])$ is compactly embedded in $\mathfrak{C}([0,1])$.

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 3.30 we know that the minimizer satisfies

$$\mathscr{B}^* \mathscr{B} z = \mathscr{B}^* w,$$

which is equivalent to

$$(\mathscr{A}^* \mathscr{A} + I) \, z = \mathscr{A}^* y + z_0.$$

Since $(\mathscr{A}^* \mathscr{A} + I)$ is injective[18], Lemma 4.16 shows that it is continuously invertible, i.e. $\left\|(\mathscr{A}^* \mathscr{A} + I)^{-1}\right\| < \infty$. Hence,

$$\|z\|_{\mathbb{X}} = \left\|(\mathscr{A}^* \mathscr{A} + I)^{-1} (\mathscr{A}^* y + z_0)\right\| \leqslant \left\|(\mathscr{A}^* \mathscr{A} + I)^{-1}\right\| (\|\mathscr{A}^*\| \, \|y\|_{\mathbb{X}} + \beta \, \|z_0\|_{\mathbb{X}}),$$

that is, the solution x of the Tikhonov regularization (4.9) is not only unique but also depends continuously on the data y, and this concludes the proof. ∎

4.2. Application of adjoint in the wellposedness of linear operator equation. In this section we are interested in the well-posedness (in the sense of Hadamard in Definition 4.13) of operator equation $\mathscr{A} x = y$, where $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ is linear and continuous. Our exposition follows [29] closely. We begin with a key result [6, 1, 29].

Lemma 4.18. *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be linear and continuous. Then*

$$\mathscr{A} \text{ is bounded below } \Leftrightarrow \exists \alpha > 0 : \|\mathscr{A} u\|_{\mathbb{Y}} \geqslant \alpha \, \|u\|_{\mathbb{X}} \Leftrightarrow \begin{cases} \mathscr{A} \text{ is injective,} \\ \mathsf{R}(\mathscr{A}) \text{ is closed.} \end{cases}$$

Proof. We provide a proof in Proof 43. ∎

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 4.19. *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

$$\|\mathscr{A} u\|_{\mathbb{Y}} \geqslant \alpha \, \|u\|_{\mathbb{X}}, \quad \forall u \in \mathbb{X}.$$

4) There exists $\alpha > 0$ such that

$$\inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A} u, v)_{\mathbb{Y}}}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geqslant \alpha.$$

---

[18] 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 $x = \theta$.

Proof. We only need to show 1) $\Leftrightarrow$ 2) as the equivalence between 2) and 3) is due to Lemma 4.18, and 4) is simply a restatement of 3). We have

$$\mathscr{A}^* \text{ is surjective}$$
$$\Updownarrow$$
$$\mathsf{R}(\mathscr{A}^*) = \mathbb{X} \text{ and thus } \mathsf{R}(\mathscr{A}^*) \text{ closed}$$
$$\Updownarrow \qquad\qquad \text{The closed range Theorem 3.13}$$
$$\mathsf{N}(\mathscr{A}) = \mathsf{R}(\mathscr{A}^*)^{\perp} = \{\theta\} \text{ and } \mathsf{R}(\mathscr{A}) \text{ closed}$$
$$\Updownarrow$$
$$\mathscr{A} \text{ is injective and } \mathsf{R}(\mathscr{A}) \text{ closed.}$$

$\blacksquare$

The following twin counterpart of Theorem 4.19 characterizes the surjectivity of $\mathscr{A}$ via the adjoint $\mathscr{A}^*$.

Theorem 4.20. *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*

$$\|\mathscr{A}^* v\|_{\mathbb{X}} \geqslant \alpha \|v\|_{\mathbb{Y}}, \quad \forall v \in \mathbb{Y}.$$

4) *There exists $\alpha > 0$ such that*
$$\inf_{v \in \mathbb{Y}} \sup_{u \in \mathbb{X}} \frac{(u, \mathscr{A}^* v)_{\mathbb{X}}}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geqslant \alpha.$$

Combining Theorem 4.19 and Theorem 4.20 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 4.21. *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 $\inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A} u, v)_{\mathbb{Y}}}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geqslant \alpha$, 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 4.19, 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 4.20. $\blacksquare$

Example 4.22. Consider $\mathcal{A} : \mathbb{R}^n \mapsto \mathbb{R}^m$. We are interested in applying Lemma 4.21 to find conditions for the linear system of equations $\mathcal{A}\boldsymbol{u} = \boldsymbol{y}$ to have a unique solution. To that end, we suppose $\mathcal{A}$ is bijective. Thus, both $\mathcal{A}$ and $\mathcal{A}^*$ are injective. By the rank-nullity theorem (3.12) we have

$$n = dim(\mathsf{N}(\mathcal{A})) + dim(\mathsf{R}(\mathcal{A})).$$

Since A is injective, we have $dim(\mathsf{N}(\mathcal{A})) = 0$, and hence

$$n = dim(\mathsf{R}(\mathcal{A})) \leqslant m.$$

Following similar arguments for the injectivity of $\mathcal{A}^*$, we have:

$$m = dim(\mathbb{R}(\mathcal{A}^*)) \leqslant n.$$

Therefore, it is necessary that $n = m$ for the bijectivity of $\mathcal{A}$. Further, the inf-sup condition in Lemma 4.21 says:

$$0 < \alpha \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^n} \sup_{\boldsymbol{v} \in \mathbb{R}^n} \frac{(\mathcal{A}\boldsymbol{u}, \boldsymbol{v})}{\|\boldsymbol{u}\|_{\mathbb{R}^n} \|\boldsymbol{v}\|_{\mathbb{R}^n}} \leqslant \inf_{\boldsymbol{u} \in \mathbb{R}^n} \frac{\|\mathcal{A}\boldsymbol{u}\|_{\mathbb{R}^n}}{\|\boldsymbol{u}\|_{\mathbb{R}^n}} = \sigma_{min}(\mathcal{A}).$$

$$\implies 0 < \alpha \leqslant \sigma_{min}(\mathcal{A}),$$

where $\sigma_{min}(\mathcal{A})$ denotes the smallest singular value of $\mathcal{A}$. We conclude the necessary and sufficient for a linear system of equations $\mathcal{A}\boldsymbol{u} = \boldsymbol{y}$ to have a unique solution is that the matrix $\mathcal{A}$ is square and invertible. This is consistent with what we know from linear algebra.

Theorem 4.23 (Banach-Nečas-Babuška). *Let $\mathscr{A} : \mathbb{X} \to \mathbb{Y}$ be the unique linear and continuous operator (see Footnote 12) 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 $|a(u, v)| \leqslant \beta \|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}$ 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 $\inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{a(u,v)}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geqslant \alpha$, 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:

$$\|u\|_{\mathbb{Y}} \leqslant \frac{1}{\alpha} \|y\|_{\mathbb{Y}}.$$

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 4.21, 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 $\mathscr{A}$:

$$\alpha \|u\|_{\mathbb{X}} \leqslant \|\mathscr{A}u\|_{\mathbb{Y}} = \sup_{v \in \mathbb{Y}} \frac{(\mathscr{A}u, v)_{\mathbb{Y}}}{\|v\|_{\mathbb{Y}}} = \sup_{v \in \mathbb{Y}} \frac{a(u, v)_{\mathbb{Y}}}{\|v\|_{\mathbb{Y}}} = \sup_{v \in \mathbb{Y}} \frac{(y, v)_{\mathbb{Y}}}{\|v\|_{\mathbb{Y}}} = \|y\|_{\mathbb{Y}}. \quad \blacksquare$$

Remark 4.24. The condition $\inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{a(u,v)}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geqslant \alpha$, 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 4.25. Consider solving the diferential equation $u' := \frac{du}{dx} = f$ 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 $u \in \mathbb{H}_0^1(0, 1) := \{u \in \mathbb{L}^2(0, 1), \ u' \in \mathbb{L}^2(0, 1), \ u(0) = 0\}$ such that:

$$\left(u', v\right)_{\mathbb{L}^2} = (f, v)_{\mathbb{L}^2}, \quad \forall v \in \mathbb{L}^2(0, 1).$$

We choose $\mathbb{X} = \mathbb{H}_0^1(0, 1)$, $\mathbb{Y} = \mathbb{L}^2(0, 1)$ and $\mathbb{F} = \mathbb{R}$, and are going to use Theorem 4.23 to show that the differential equation is well-posed.

The continuity of the bilinear form $a(u, v)$ is clear as

$$|a(u, v)| = \left|(u', v)\right| \leqslant \left\|u'\right\|_{\mathbb{L}^2} \|v\|_{\mathbb{L}^2} \leqslant \|u\|_{\mathbb{H}^1} \|v\|_{\mathbb{L}^2}.$$

We next verify the inf-sup condition Item **C1)**. By a simple integration and the Cauchy-Schwarz inequality (see also Example 4.35) we obtain the following the Poincaré-Friedrichs inequality

$$\left\|u'\right\|_{\mathbb{L}^2} \geqslant \|u\|_{\mathbb{L}^2},$$

and thus the inf-sup condition holds since

$$\inf_{u \in \mathbb{H}^1} \sup_{v \in \mathbb{L}^2} \frac{a(u, v)}{\|u\|_{\mathbb{H}^1} \|v\|_{\mathbb{L}^2}} = \inf_{u \in \mathbb{H}^1} \frac{\|u'\|_{\mathbb{L}^2}}{\|u\|_{\mathbb{H}^1}} \geqslant \frac{1}{\sqrt{2}}.$$

Now we verify the injectivity of the adjoint, i.e Item **C2)**. We start from

$$(4.11) \qquad \qquad \left(u', v\right)_{\mathbb{L}^2} = 0, \quad \forall u \in \mathbb{H}_0^1(0, 1).$$

Since $\mathfrak{C}_0^\infty(0, 1) \subset \mathbb{H}_0^1(0, 1)$, we have:

$$\left(\psi', v\right)_{\mathbb{L}^2} = 0, \quad \forall \psi \in \mathfrak{C}_0^\infty(0, 1).$$

By definition of the distributional derivative we arrive at

$$\left\langle \psi, v' \right\rangle_{\mathbb{L}^2} = 0, \quad \forall \psi \in \mathfrak{C}_0^\infty(0, 1),$$

which implies that v is a constant function. Now in (4.11) 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 4.26 (Friedrichs' system). We consider the abstract problem $\mathscr{A}u = y$ where $\mathscr{A}$ is the Friedrichs' operator defined in Example 4.5. We choose $\mathbb{Y} = \mathbb{L}^2(\Omega)$ and

$$\mathbb{X} = \{x \in \mathbb{H}_\mathscr{A} : (\mathcal{D} - \mathcal{M}) x = 0 \text{ on } \partial\Omega\},$$

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 [29, Theorem 5.7]), and thus applying Theorem 4.23 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 [33, 30, 45, 31].

When $\mathbb{Y} = \mathbb{X}$ and the sequilinear form is symmetric, the inf-sup condition is both necessary and sufficient for bijectivity.

Lemma 4.27. *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*

$$\exists \alpha > 0 \ \text{such that} \ \inf_{u \in \mathbb{X}} \sup_{v \in \mathbb{Y}} \frac{a\,(u,v)}{\|u\|_{\mathbb{X}} \|v\|_{\mathbb{Y}}} \geqslant \alpha.$$

Proof. We need to prove only the second condition in the second statement of Theorem 4.23, namely the injectivity of the adjoint in Item **C2)**. But this is obvious due to symmetry:

$$0 = \overline{a\,(w,v)} = a\,(v,w)\,, \quad \forall w \in \mathbb{X} \implies 0 = \sup_{w \in \mathbb{X}} \frac{a\,(v,w)}{\|w\|_{\mathbb{X}}} \geqslant \alpha \|v\|_{\mathbb{X}} \implies v = \theta. \qquad \blacksquare$$

Remark 4.28. 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 4.23. 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 4.29 (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*

(Coercivity) $$a\,(v,v) \geqslant \alpha \|v\|_{\mathbb{X}}^2\,, \quad \forall v \in \mathbb{X},$$

then there exists a unique solution u and $\|u\| \leqslant \frac{1}{\alpha} \|y\|_{\mathbb{X}}$.

Proof. We need to verify the two conditions in the second statement of Theorem 4.23. The inf-sup condition in Item **C1)** is clear from (Coercivity) as

$$\alpha \|v\|_{\mathbb{X}} \leqslant \frac{a\,(v,v)}{\|v\|_{\mathbb{X}}} \leqslant \sup_{w \in \mathbb{X}} \frac{a\,(v,w)}{\|w\|_{\mathbb{X}}}.$$

For the injectivity of the adjoint in Item **C2)**, we note that

$$a\,(w,v) = 0\,, \quad \forall w \in \mathbb{X} \implies 0 = \sup_{w \in \mathbb{X}} a\,(w,v) \geqslant a\,(v,v) \geqslant \alpha \|v\|_{\mathbb{X}}^2 \implies v = \theta. \qquad \blacksquare$$

When the sequilinear form is symmetric and positive, it turns out that (Coercivity) is both sufficient and necessary.

Corollary 4.30. *Consider the variational equation* $a\left(u,v\right) = \left(y,v\right)_{\mathbb{X}}$, $\forall v \in \mathbb{X}$, *with a continuous sequilinear form* $a : \mathbb{X} \times \mathbb{X} \to \mathbb{F}$, *with continuity constant* β. *Suppose* $a\left(\cdot,\cdot\right)$ *is symmetric, i.e.,* $a\left(w,v\right) = \overline{a\left(v,w\right)}$ *and positive, i.e.,* $a\left(v,v\right) > 0, \forall v \neq \theta$. *Then, there exists a unique solution* u *iff the coercivity condition* (Coercivity) *holds.*

Proof. Lemma 4.29 proves the sufficiency, and we need to show the necessity. Since the sequilinear form $a\left(\cdot,\cdot\right)$ is symmetric and positive, it defines an inner product in $\mathbb{X}$ and the induced norm is

$$\|v\|_a := \sqrt{a\left(v,v\right)}.$$

Thus, by the inf-sup condition, the Cauchy-Schwarz inequality, and the continuity of $a\left(\cdot,\cdot\right)$, we have

$$\alpha \|v\|_{\mathbb{X}} \leqslant \sup_{w \in \mathbb{X}} \frac{a\left(w,v\right)}{\|w\|_{\mathbb{X}}} \leqslant \sup_{w \in \mathbb{X}} \frac{\sqrt{a\left(v,v\right)}\sqrt{a\left(w,w\right)}}{\|w\|_{\mathbb{X}}} \leqslant \sqrt{\beta}\sqrt{a\left(v,v\right)},$$

and this ends the proof. ∎

Example 4.31. We now consider the variational equation $a\left(u,v\right) = \left(y,v\right)_{\mathbb{Y}}$ where the bilinear form $a\left(u,v\right)$, the spaces $\mathbb{X}$ and $\mathbb{Y}$, and other specifications are described in Example 4.3. We have shown that $a\left(u,v\right)$ is symmetric and continuous on $\mathbb{H}_0^1\left(\Omega\right)$. It is clearly positive if we assume that z is bounded away from $-\infty$. What remains is to show that $a\left(u,v\right)$ is coercive. Recall the Poincaré-Friedrichs inequality (see, e.g., [29, 6, 16]) for $\mathbb{H}_0^1\left(\Omega\right)$: there exists a constant c depending on only Ω such that

$$c \|u\|_{\mathbb{L}^2} \leqslant \|\nabla u\|_{\mathbb{L}^2}, \quad \forall u \in \mathbb{H}_0^1\left(\Omega\right).$$

It follows that for any $v \in \mathbb{H}_0^1\left(\Omega\right)$ we have

$$a\left(v,v\right) = \left(e^z \nabla v, \nabla v\right)_{\mathbb{L}^2}^2 \geqslant e^{\inf z} \|\nabla v\|_{\mathbb{L}^2}^2 \geqslant \min\left\{1,c\right\} \frac{1}{2} e^{\inf z} \|v\|_{\mathbb{H}^1}^2,$$

and this ends the proof.

4.3. Application of adjoint to the Sturm-Liouville problem and generalized Fourier series. In this section, we are interested in self-adjoint operator $\mathscr{A}$ in infinite dimensions. Our exposition mostly follows [71]. To the end of this section, for any $\mathscr{A} \in \mathfrak{L}\left(\mathbb{X},\mathbb{Y}\right)$ we assume that its domain $\mathsf{D}\left(\mathscr{A}\right) \subseteq \mathbb{U}$ is dense in $\mathbb{X}$.

Definition 4.32 (Closed linear operators). *A linear operator* $\mathscr{A} \in \mathfrak{L}\left(\mathbb{X},\mathbb{Y}\right)$ *is called closed iff its graph*

$$\mathcal{G}_{\mathscr{A}} := \left\{[u, \mathscr{A}u] : u \in \mathscr{D}\left(\mathscr{A}\right)\right\}$$

is closed.

Clearly, any continuous linear operator is necessarily closed. The following are standard results for closed operators [71].

Lemma 4.33. *The following hold:*
- *If* $\mathscr{A}$ *is closed, so is* $\mathscr{A}^*$.
- $\mathscr{A}$ *is closed and* $\mathsf{D}\left(\mathscr{A}\right) = \mathbb{U}$ *iff* $\mathscr{A} \in \mathscr{B}\left(\mathbb{U},\mathbb{V}\right)$.
- *If* $\mathsf{D}\left(\mathscr{A}\right) = \mathbb{U}$, *then* $\mathscr{A}^*$ *is continuous, and hence* $\mathsf{D}\left(\mathscr{A}^*\right)$ *is closed.*

- *If $\mathscr{A}$ is closed, then $\mathsf{D}\left(\mathscr{A}^*\right)$ is dense in $\mathbb{V}$.*

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 $\mathbb{V} \xhookrightarrow[\text{dense}]{\text{compact}} \mathbb{X}$. Let $\mathscr{A}$ be a linear operator $\mathscr{A} : \mathsf{D}\left(\mathscr{A}\right) \subset \mathbb{V} \xhookrightarrow[\text{dense}]{\text{compact}} \mathbb{X} \to \mathbb{X}$ with the domain defined as

$$\mathsf{D}\left(\mathscr{A}\right) := \{x \in \mathbb{V} : \mathscr{A}x \in \mathbb{X}\}.$$

We also assume that the associate sequilinear form $a\left(u,v\right) := \left(\mathscr{A}u, v\right)_{\mathbb{X}}$ is defined for $u \in \mathsf{D}\left(\mathscr{A}\right)$ and $v \in \mathbb{V}$, and is continuous on $\mathbb{V}$. Furthermore, we assume that the $a\left(u,v\right)$ is $\mathbb{V}$-elliptic in the following sense: there exists $c > 0$ such that

$$a\left(v,v\right) + \overline{a\left(v,v\right)} \geqslant 2c \left\|v\right\|_{\mathbb{V}}^2.$$

Finally, we assume that $a\left(\cdot,\cdot\right)$ is symmetric, i.e.,

$$a\left(u,v\right) = \overline{a\left(v,u\right)}, \quad \forall u, v \in \mathbb{V},$$

which is, again, equivalent to the self-adjointness of $\mathscr{A}$ on $\mathbb{V}$. Recall that Lemma 3.21 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 4.34. *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 $\left\|v\right\|_{\mathbb{X}} = 1$,*
- *$\left(v_n, v_m\right)_{\mathbb{X}} = 0$ for all $n \neq m$,*
- *$0 < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n \to \infty$ when $n \to \infty$, and*
- *$\{v_n\}_{n=1}^{\infty}$ is a basis of $\mathbb{X}$.*

Proof. We sketch the proof and more details can be found in [72]. The keys are: i) the $\mathbb{V}$-elliptic condition implies coercivity of $\mathscr{A}$ on $\mathbb{V}$. By Lax-Milgram Lemma 4.29, $\mathscr{A} : \mathbb{V} \to \mathbb{V}' \equiv \mathbb{V}$ is a continuous bijection. The restriction of $\mathscr{A}$ on $\mathsf{D}\left(\mathscr{A}\right)$ is thus surjective on $\mathbb{X}$, and $\mathscr{A}^{-1} : \mathbb{X} \to \mathsf{D}\left(\mathscr{A}\right) \subset \mathbb{V}$ 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 4.9 then ensures the existence of the eigenpairs μ_n, v_n of $\mathscr{A}^{-1}$, where $\{v_n\}_n$ is a basis of $\mathsf{D}\left(\mathscr{A}\right)$, 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 closed operator and this leads to the conclusion that $\{v_n\}_{n=1}^{\infty}$ is a also basis of $\mathbb{X}$. ∎

Theorem 4.34 provides a generalized Fourier series in $\mathbb{X}$. In particular, for any function $f \in \mathbb{X}$, we have

$$f = \sum_{n=1}^{\infty} \left(f, v_n\right)_{\mathbb{X}} v_n,$$

where equality means convergence in the topology generated by the $\mathbb{X}$-norm. Below are a few special cases that lead to the standard Fourier series and a general Sturm-Liouville problem.

Example 4.35. Consider the following Sturm-Liouville problem: seek λ and v such that

$$(4.12) \qquad \begin{cases} -\frac{d^2 v}{dx^2} = \lambda v & \text{in } \Omega = (0,1), \\ v = 0 & \text{on } \partial\Omega = \{0,1\}, \end{cases}$$

where the derivative is understood in the classical sense sense. We solve this problem using Theorem 4.34. To that end, we define the symmetric sequilinear form as

$$a(u,v) := \left(\frac{du}{dx}, \frac{dv}{dx} \right)_{\mathbb{L}^2(\Omega)} = \int_0^1 \frac{du}{dx} \frac{\overline{dv}}{dx} \, dx,$$

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 $\mathbb{V} \xrightarrow[\text{dense}]{\text{compact}} \mathbb{X} := \mathbb{L}^2(0,1)$ [71], 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: $\forall v \in \mathfrak{C}_0^1[0,1]$

$$\left\| \frac{dv}{dx} \right\|_{\mathbb{L}^2(0,1)} \geqslant \|v\|_{\mathbb{L}^2(0,1)},$$

which also holds for any $v \in \mathbb{H}_0^1(0,1)$ due to the density of $\mathfrak{C}_0^1(0,1)$ in $\mathbb{V}$. This leads to the $\mathbb{V}$-ellipticity of $a(u,v)$ as

$$a(v,v) + \overline{a(v,v)} \geqslant \|v\|_{\mathbb{H}_0^1(0,1)}^2.$$

Thus, Theorem 4.34 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 $a(\cdot,\cdot)$,

$$\begin{array}{rcll} (\mathscr{A}w,v)_{\mathbb{L}^2(0,1)} & = & \int_0^1 \frac{dw}{dx}\frac{\overline{dv}}{dx}\,dx & \forall v \in \mathbb{H}_0^1(0,1) \\ & \Updownarrow & & \text{density of } \mathfrak{C}_0^\infty(0,1) \text{ in } \mathbb{H}_0^1(0,1) \\ (\mathscr{A}w,\varphi)_{\mathbb{L}^2(0,1)} & = & \int_0^1 \frac{dw}{dx}\frac{\overline{d\varphi}}{dx}\,dx & \forall \varphi \in \mathfrak{C}_0^\infty(0,1) \\ & \Updownarrow & & \text{definition of distributional derivative} \\ (\mathscr{A}w,\varphi)_{\mathbb{L}^2(0,1)} & = & \left\langle -\frac{d^2 w}{dx^2}, \varphi \right\rangle & \forall \varphi \in \mathfrak{C}_0^\infty(0,1) \\ & \Updownarrow & & \\ \mathscr{A}w & = & -\frac{d^2 w}{dx^2} & \text{in } \mathbb{L}^2(0,1) \end{array}$$

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,

$$(4.13) \qquad -\frac{d^2 v_n}{dx^2} = \lambda_n v_n, \text{ in } \mathbb{L}^2(0,1),$$

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 $\mathfrak{C}_0(0,1)$ (see, e.g., [16, Theorem 8.2]) the eigenvalue problem (4.13) holds in the classical sense, which is exactly the Sturm-Liouville problem (4.12). Thus, eigenfunctions of the Sturm-Liouville problem (4.12) 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 4.36. *Let $\mathbb{V}$ and $\mathbb{X}$ be given in Theorem 4.34. Suppose the sequilinear form $a(\cdot,\cdot)$ is symmetric and continuous on $\mathbb{V}$. Assume that there exists $c > 0$ such that for some $\lambda \in \mathbb{R}$*

$$a(v,v) + \overline{a(v,v)} + 2\lambda \|v\|_{\mathbb{X}}^2 \geqslant 2c \|v\|_{\mathbb{V}}^2, \quad \forall v \in \mathbb{V}.$$

Then, there exists an orthonormal sequence of eigenfunctions of $\mathscr{A}$, which is a basis for $\mathbb{X}$ and the corresponding eigenvalues satisfies $-\lambda < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n \to \infty$ 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

$$b(v,v) + \overline{b(v,v)} \geqslant 2c \|v\|_{\mathbb{V}}^2, \quad \forall v \in \mathbb{V},$$

i.e., $b(u,v)$ is $\mathbb{V}$-elliptic. Theorem 4.34 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 $0 < \gamma_1 \leqslant \gamma_2 \leqslant \ldots \leqslant \gamma_n \to \infty$ when $n \to \infty$. As a result, $\{\lambda_n, v_n\}_{n=1}^{\infty}$ with $\lambda_n := \gamma_n - \lambda$, and thus $-\lambda < \lambda_1 \leqslant \lambda_2 \leqslant \ldots \leqslant \lambda_n \to \infty$ when $n \to \infty$, and $\{v_n\}_{n=1}^{\infty}$ being a basis of $\mathbb{X}$, and this concludes the proof. ∎

Remark 4.37. We can use Corollary 4.36 for Example 4.35 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 4.38. Consider the following Sturm-Liouville problem: seek κ and v such that

$$(4.14) \qquad \begin{cases} -\frac{d^2 v}{dx^2} = \kappa v & \text{in } \Omega = (0,1), \\ \frac{dv}{dx} = 0 & \text{on } \partial\Omega = \{0,1\}, \end{cases}$$

where the derivative is understood in the classical sense. The sequilinear $a(\cdot,\cdot)$ is defined the same as in Example 4.35 with $\mathbb{V} := \mathbb{H}^1(0,1)$ and $\mathbb{X} := \mathbb{L}^2(0,1)$. We still have $\mathbb{V} \xhookrightarrow[\text{dense}]{\text{compact}} \mathbb{X} := \mathbb{L}^2(0,1)$ [71]. 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 $u \in \mathsf{D}(\mathscr{A}) := \{w \in \mathbb{V} : \mathscr{A}w \in \mathbb{X} \text{ and } \frac{dw}{dx} = 0 \text{ on } \partial\Omega = \{0,1\}\}$, and $v \in \mathbb{V}$. Note that unlike Example 4.35 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 (4.14). The continuity and symmetry of $a(\cdot,\cdot)$ on $\mathbb{V}$ are clear. Furthermore, $a(\cdot,\cdot)$ satisfies Corollary 4.36 for any $\lambda > 0$ and $c = \min\{1, \lambda\}$. Using a similar argument as in Example 4.35, the eigenpairs λ_n, v_n of $\mathscr{A}$ are exactly the solutions of (4.14), and in particular $v_n = \sqrt{2}\cos(n\pi x)$ for $n \geqslant 1$ and $v_0 = 1$. By Corollary 4.36, $\{v_n\}_{n=0}^{\infty}$ is another Fourier basis (*cosine series*) for $\mathbb{L}^2(0,1)$.

Example 4.39 (A more general Sturm-Liouville problem). We now generalize Example 4.35 and Example 4.38: seek κ and v such that

$$
(4.15) \qquad \begin{cases} \frac{1}{\rho} \left[\frac{d}{dx} \left(p \frac{dv}{dx} \right) + qv \right] = \kappa v & \text{in } \Omega = (0,1), \\ \alpha v(0) + \beta \frac{dv}{dx}(0) = 0, \qquad \gamma v(1) + \delta \frac{dv}{dx}(1) = 0, \end{cases}
$$

where $\rho \in \mathfrak{C}[0,1]$ and $\rho > 0$, $p \in \mathfrak{C}^1[0,1]$ and $p < 0$, and $q \in \mathfrak{C}[0,1]$. 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 $u \in \mathsf{D}(\mathscr{A}) := \left\{ w \in \mathbb{H}^1_\rho(0,1) : \mathscr{A}w \in \mathbb{X}, \alpha w(0) + \beta \frac{dw}{dx}(0) = 0 \text{ and } \gamma w(1) + \delta \frac{dw}{dx}(1) = 0 \right\}$, and $v \in \mathbb{V}$. By a similar distributional argument as in Example 4.35, one can show that $\mathscr{A}w = \frac{1}{\rho} \left[\frac{d}{dx} \left(p \frac{dw}{dx} \right) + qw \right]$ in $\mathbb{L}^2(0,1)$. Furthermore, it is clear that $a(\cdot,\cdot)$ satisfies Corollary 4.36 for any $\lambda > -q$ and $c = \min \left\{ \left\| \frac{\rho}{p} \right\|_\infty^{-1}, \left\| \frac{\rho}{\lambda+q} \right\|_\infty^{-1} \right\}$. We thus conclude that the eigenpairs λ_n and $v_n \in \mathsf{D}(\mathscr{A})$ satisfy

$$
(4.16) \qquad \frac{1}{\rho} \left[\frac{d}{dx} \left(p \frac{dv_n}{dx} \right) + qv_n \right] = \lambda_n v_n,
$$

which is equivalent to

$$
v_n = \int \frac{1}{p} \int \rho \left(\lambda_n - q \right) v_n \, dx \, dy,
$$

which, together with the embedding result in [16, Theorem 8.2], shows that the eigenvalue problem (4.16) holds in the classical sense, which is exactly the Sturm-Liouville problem (4.15). Thus, eigenfunctions of the Sturm-Liouville problem (4.15) 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 4.40. Next, we consider the following Sturm-Liouville problem that is not covered by Example 4.39: seek κ and v such that

$$
(4.17) \qquad \begin{cases} -\frac{d^2 v}{dx^2} = \kappa v & \text{in } \Omega = (0,1), \\ v(0) = v(1) & \text{and } \frac{dv}{dx}(0) = \frac{dv}{dx}(1), \end{cases}
$$

where the derivative is understood in the classical sense. The sequilinear $a(\cdot,\cdot)$ is defined the same as in Example 4.35 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 $u \in \mathsf{D}(\mathscr{A}) := \left\{ w \in \mathbb{V} : \mathscr{A}w \in \mathbb{X}, v(0) = v(1) \text{ and } \frac{dv}{dx}(0) = \frac{dv}{dx}(1) \right\}$, and $v \in \mathbb{V}$. Similar to Example 4.38,

$a\left(\cdot,\cdot\right)$ is obviously symmetric and continuous on $\mathbb{V}$. Furthermore, $a\left(\cdot,\cdot\right)$ satisfies Corollary 4.36 for any $\lambda > 0$ and $c = \min\left\{1,\lambda\right\}$. Using a similar distributional argument as in Example 4.35, the eigenpairs λ_n, v_n of $\mathscr{A}$ are exactly the solutions of (4.17), and in particular $v_0 = 1$ and $v_n \in \left\{\sqrt{2}\cos\left(2n\pi x\right), \sqrt{2}\sin\left(2n\pi x\right)\right\}$ for $n \geqslant 1$, which is the usual Fourier basis for $\mathbb{L}^2\left(0,1\right)$.

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., [6]).

4.4. Application of adjoint in PDE-constrained optimization. We have seen the important role of adjoint in constrained optimization in subsection 3.5, especially constrained optimization with equality constraints that have separable structure (see Corollary 3.56 and Lemma 3.57). We have also seen how the adjoint looks like and how it helps compute the gradient of deep neural networks (DNN) efficiently as backpropagation in subsection 3.6. In this section, we shall work out the details of adjoint equation and the reduced gradient for optimization problems constrained by partial differential equations (PDE-constrained optimization). We consider prototype steady state (time-independent) PDEs of elliptic and hyperbolic types. The goal is to show we translate the abstract results in Lemma 3.57 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 [57, 12, 11, 13, 52].

Example 4.41 (Advection-PDE-constrained optimization problem). Consider the following PDE-constrained optimization problem

$$\min_{z,u} \frac{1}{2} \int_\Omega u^2 \, d\Omega$$

subject   to

(4.18a) $$\boldsymbol{\beta} \cdot \nabla u = 0, \quad \text{in } \Omega,$$

(4.18b) $$\boldsymbol{\beta} \cdot \boldsymbol{n}u = z, \quad \text{in } \partial\Omega_{\text{in}},$$

where $u \in \mathbb{H}^1_{\boldsymbol{\beta}}\left(\Omega\right) := \left\{u : u \in \mathbb{L}^2\left(\Omega\right) \text{ and } \boldsymbol{\beta} \cdot \nabla u \in \mathbb{L}^2\left(\Omega\right)\right\}$. See Example 4.4 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 3.57. Note that for $u \in \mathbb{H}^1_{\boldsymbol{\beta}}\left(\Omega\right)$, its trace (in fact weighted trace with weight $\left|\boldsymbol{\beta} \cdot \boldsymbol{n}\right|$) on $\partial\Omega_{\text{in}}$ belongs to $\mathbb{L}^2\left(\partial\Omega_{\text{in}}\right)$. The correct space for z is thus $\mathbb{L}^2\left(\partial\Omega_{\text{in}}\right)$ with a weighted norm (see Footnote 20). Since the constraints map $[u,z] \in \mathbb{X} \times \mathbb{Z} := \mathbb{H}^1_{\boldsymbol{\beta}}\left(\Omega\right) \times \mathbb{L}^2\left(\partial\Omega_{\text{in}}\right)$ to $\mathbb{Y} := \mathbb{L}^2\left(\Omega\right) \times \mathbb{L}^2\left(\partial\Omega_{\text{in}}\right)$, the Lagrange multiplier $y = [v,w]$ has two components $v \in \mathbb{L}^2\left(\Omega\right)$ and $w \in \mathbb{L}^2\left(\partial\Omega_{\text{in}}\right)$, respectively. Our task is to find the explicit form of the first order optimality condition (3.22) which, we recall, is a special case of the first order optimality condition via

the Lagrangian multiplier Theorem 3.52. 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 3.52 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 (3.17), *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 (3.17) in this case reads: $\forall [h, r] \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) \times \mathbb{L}^2(\partial\Omega_{\text{in}})$,

$$\int_\Omega u h \, d\Omega - \int_\Omega (\boldsymbol{\beta} \cdot \nabla v)\, h \, d\Omega + \int_{\partial\Omega_{\text{in}}} \boldsymbol{\beta} \cdot \boldsymbol{n}\,(w + v)\, h \, ds + \int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta} \cdot \boldsymbol{n} v\, h \, ds - \int_{\partial\Omega_{\text{in}}} r\, w \, ds = 0,$$

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

$$\int_\Omega u h \, d\Omega - \int_\Omega (\boldsymbol{\beta} \cdot \nabla v)\, h \, d\Omega = 0, \quad \forall h \in \mathbb{H}^1_{\boldsymbol{\beta},0}(\Omega),$$

which implies[19]

$$-\boldsymbol{\beta} \cdot \nabla v + u = 0.$$

Consequently, the first order optimality condition reduces to: $\forall [h, r] \in \mathbb{H}^1_{\boldsymbol{\beta}}(\Omega) \times \mathbb{L}^2(\partial\Omega_{\text{in}})$,

$$(4.19) \qquad \int_{\partial\Omega_{\text{in}}} \boldsymbol{\beta} \cdot n\,(w + v)\, h \, ds + \int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta} \cdot \boldsymbol{n} v\, h \, ds - \int_{\partial\Omega_{\text{in}}} r\, w \, ds = 0,$$

which, by taking $h = 0$ on $\partial\Omega_{\text{out}}$ and $r = 0$, becomes

$$\int_{\partial\Omega_{\text{in}}} \boldsymbol{\beta} \cdot \boldsymbol{n}\,(w + v)\, h \, ds = 0,$$

---

[19] 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 [4].

which in turn gives[20]

$$w = -v \text{ on } \partial\Omega_{\text{in}},$$

that is, the adjoint variables are not independent. This can be then substituted into (4.19) to further reduce the first order optimality condition to

$$\int_{\partial\Omega_{\text{out}}} \boldsymbol{\beta} \cdot \boldsymbol{n} v \, h \, ds + \int_{\partial\Omega_{\text{in}}} r \, v \, ds = 0, \quad \forall \, [h, r] \in \mathbb{H}_{\boldsymbol{\beta}}^1(\Omega) \times \mathbb{L}^2(\partial\Omega_{\text{in}}).$$

It follows that, by taking $r = 0$ and using the surjectivity in Footnote 20, 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 (3.22c) becomes

(4.20) $$v = 0 \text{ on } \partial\Omega_{\text{in}},$$

and the adjoint equation (3.22b) reads

(4.21a) $$-\boldsymbol{\beta} \cdot \nabla v = -u \text{ in } \Omega,$$

(4.21b) $$\boldsymbol{\beta} \cdot \boldsymbol{n} v = 0 \text{ on } \partial\Omega_{\text{out}},$$

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 4.4, 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 $z = 0$. The reduced gradient can be now computed for a given z via three steps: 1) solve the *forward equation* (4.18) for $u(z)$, 2) solve the *adjoint equation* (4.21) for $v(u(z), z)$, and 3) substitute $v(u(z), z)$ into the left hand side of (4.20) to obtain the reduced gradient.

Example 4.42 (Elliptic-PDE-constrained optimization problem).

$$\min_{z,u} J(u) := \frac{1}{2} \int_{\Omega} \left(u - u^{obs} \right)^2 d\Omega$$

subject   to

(4.22a) $$-\nabla \cdot (e^z \nabla u) = 0, \quad \text{in } \Omega,$$

(4.22b) $$u = g, \quad \text{in } \partial\Omega,$$

---

[20]Note that this is true due to the fact that the trace operator $\gamma : H_{\boldsymbol{\beta},0}^1(\Omega) \rightarrow \mathbb{L}_{\boldsymbol{\beta}\cdot n}^2(\partial\Omega)$ is a continuous surjection (see, e.g., [17]), where $\mathbb{L}_{\boldsymbol{\beta}\cdot n}^2(\partial\Omega)$ is $\mathbb{L}^2(\partial\Omega)$ with the weighted inner product $(u,v)_{\mathbb{L}_{\boldsymbol{\beta}\cdot n}^2(\partial\Omega)} := \int_{\partial\Omega} |\boldsymbol{\beta} \cdot \boldsymbol{n}| \, uv \, ds.$

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 4.2: in particular, $z \in \mathfrak{C}^1(\Omega) \subset \mathbb{L}^2(\Omega)$. Here, g is the Dirichlet boundary data. This optimization is a special case of the abstract one in Lemma 3.57. Thus, the constraint maps $[u, z] \in \mathbb{X} \times \mathbb{Z} := \mathbb{H}_{\mathscr{A}} \times \mathfrak{C}^1(\Omega)$ 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 Example 4.41, we go back to Lagrangian multiplier Theorem 3.52 to derive the explicit form 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 $[u, z] \in \mathbb{H}_{\mathscr{A}} \times \mathfrak{C}^1(\Omega)$. Take an arbitrary direction $[h, r] \in \mathbb{H}_{\mathscr{A}} \times \mathfrak{C}^1(\Omega)$, the first order optimality condition (3.17), *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:

$$(4.23) \quad \int_{\Omega} \left(u - u^{obs} \right) h \, d\Omega + \int_{\Omega} \left[-\nabla \cdot (e^z \nabla v) \right] h \, d\Omega + \int_{\partial\Omega} h \, w \, ds - \int_{\partial\Omega} e^z \nabla h \cdot \boldsymbol{n} \, v \, ds$$
$$+ \int_{\partial\Omega} e^z \nabla v \cdot \boldsymbol{n} \, h \, ds + \int_{\Omega} \left[-\nabla \cdot (e^z r \nabla u) \right] v \, d\Omega = 0, \quad \forall [h, r] \in \mathbb{H}_{\mathscr{A}} \times \mathfrak{C}^1(\Omega).$$

Following a similar strategy[21] as in Example 4.41 for (4.23) gives the adjoint equation

$$(4.24a) \qquad\qquad -\nabla \cdot (e^z \nabla v) = -\left(u - u^{obs} \right), \quad \text{in } \Omega,$$

$$(4.24b) \qquad\qquad\qquad v = 0, \quad \text{in } \partial\Omega.$$

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 4.2, which is not a surprise. The first order optimality condition is thus reduced to: $\quad \forall [h, r] \in \mathbb{H}_{\mathscr{A}} \times \mathfrak{C}^1(\Omega),$

$$(4.25) \qquad\qquad \int_{\partial\Omega} h \, w \, ds + \int_{\partial\Omega} e^z \nabla v \cdot \boldsymbol{n} \, h \, ds + \int_{\Omega} \left[-\nabla \cdot (e^z r \nabla u) \right] v \, d\Omega = 0,$$

which, by taking $r = 0$, gives,

$$w = -e^z \nabla v \cdot \boldsymbol{n}.$$

---

[21] Here we take $h \in \mathfrak{C}_0^{\infty}(\Omega)$ and $r = 0$ to obtain the equation (4.24a) for v. To get the boundary condition (4.24b), we then take $h \in \mathfrak{C}_0^1(\Omega)$ so that the normal trace $\nabla h \cdot \boldsymbol{n}$ is surjective on $\mathbb{L}^2(\partial\Omega)$.

Similar to Example 4.41, 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

$$\int_\Omega \left[-\nabla \cdot (e^z r \nabla u) \right] v \, d\Omega = 0, \quad \forall r \in \mathfrak{C}^1(\Omega),$$

which—after integrating by parts, using the fact that $v = 0$ on $\partial\Omega$, and using the fact that $\mathfrak{C}^1(\Omega)$ is dense in $L^2(\Omega)$—gives the control equation

$$(4.26) \qquad\qquad\qquad\qquad e^z \nabla u \cdot \nabla v = 0.$$

The reduced gradient can be now computed for a given z via three steps: 1) solve the *forward equation* (4.22) for $u(z)$, 2) solve the *adjoint equation* (4.24) for $v(u(z), z)$, and 3) substitute $u(z)$ and $v(u(z), z)$ into the left hand side of (4.26) to obtain the reduced gradient.

5. Conclusions. This paper has put together in one place the roles of adjoint in various disciplines of mathematics, sciences, and engineering. The objective is to systematically compile these materials on the same mathematical footing starting from the basic definitions. We aim to provide a unified perspective and understanding of adjoint. This work could give broader views and better insights into the application of adjoint beyond a single community. We have established general results and then specified them to each application with sufficient details including the connections among them. This paper is written as an interdisciplinary tutorial on adjoint with results and applications in both finite-dimensional and infinite-dimensional Hilbert spaces. We have shown how adjoint can solve problems and facilitate progress in various fields through specific examples including linear algebra (e.g. eigendecomposition and the singular value decomposition), ordinary differential equations (an epidemic model), partial differential equations (of elliptic, hyperbolic, and Friedrichs' types), neural networks (feed-forward deep neural networks), least squares and inverse problems (with Tikhonov regularization), and PDE-constrained optimization.

Appendix A. Proofs of some results in the paper.

Proof for Example 4.2. We next find $\mathscr{A}^*$ and $\mathsf{D}(\mathscr{A}^*)$. Take any $v \in \mathsf{D}(\mathscr{A}^*)$, by Defini-

tion 4.1 we have

$$(\mathscr{A}u, v)_{\mathbb{L}^2(\Omega)} = (u, \mathscr{A}^*v)_{\mathbb{L}^2(\Omega)}, \qquad \forall u \in \mathsf{D}(\mathscr{A})$$

$$\Updownarrow \qquad\qquad\qquad \forall \varphi \in \mathfrak{C}_0^\infty(\Omega) \text{ dense in } \mathsf{D}(\mathscr{A})$$

$$(\mathscr{A}\varphi, v)_{\mathbb{L}^2(\Omega)} = (\varphi, \mathscr{A}^*v)_{\mathbb{L}^2(\Omega)}$$

$$\Updownarrow \qquad\qquad\qquad \text{definition of distributional derivative}$$

$$\langle \varphi, -\nabla \cdot (e^z \nabla v)\rangle = (\varphi, \mathscr{A}^*v)_{\mathbb{L}^2(\Omega)}$$

$$\Updownarrow$$

$$-\nabla \cdot (e^z \nabla v) = \mathscr{A}^*v \in \mathbb{L}^2(\Omega)$$

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 for Example 4.5. We now derive the adjoint $\mathscr{A}^*$ (and its domain) of the Friedrichs' operator in Example 4.5. Proceed as in Proof 41, we can easily see that

$$\mathscr{A}^*v = -\sum_{k=1}^n \partial_k (\mathcal{A}_k v) + \mathcal{C}^T v \in \mathbb{L}^2(\Omega),$$

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

$$(u, \mathcal{D}v)_{\mathbb{L}^2(\partial\Omega)} = 0, \quad \forall u \in \mathsf{D}(\mathscr{A}) \text{ and } v \in \mathsf{D}(\mathscr{A}^*),$$

which in turns yields $\mathcal{D}v \in \mathsf{N}(\mathcal{D} - \mathcal{M})^\perp$ since $u \in \mathsf{N}(\mathcal{D} - \mathcal{M})$. We thus conclude

$$\mathsf{D}(\mathscr{A}^*) = \left\{ v \in \mathbb{H}_{\mathscr{A}} : \mathcal{D}v \in [\mathsf{N}(\mathcal{D} - \mathcal{M})]^\perp \right\}.$$

■

Proof of Lemma 4.18. 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 $n = n\left(\varepsilon\right)$ such that for all $i, j > n$ we have

$$\alpha \left\| x_i - x_j \right\|_{\mathbb{X}} \leqslant \left\| \mathscr{A} x_i - \mathscr{A} x_j \right\|_{\mathbb{Y}} \left\| y_i - y_j \right\|_{\mathbb{Y}} < \varepsilon$$
$$\Downarrow$$
$$\{x_i\}_{i=1}^{\infty} \text{ is Cauchy } \implies x_i \xrightarrow{\mathbb{X}} x$$
$$\Downarrow \qquad\qquad \text{continuity of } \mathscr{A}$$
$$y \xleftarrow{\mathbb{Y}} y_i = \mathscr{A} x_i \xrightarrow{\mathbb{Y}} \mathscr{A} x,$$

and thus $\mathsf{R}\left(\mathscr{A}\right)$ is closed.

For the sufficiency, $\mathsf{R}\left(\mathscr{A}\right)$ is a Banach space due to its closedness. $\mathscr{A} : \mathbb{X} \to \mathsf{R}\left(\mathscr{A}\right)$ is thus bijective, which in turn implies $\mathscr{A}^{-1}$ is linear and continuous owing to the Open Mapping Theorem [16, 63, 6, 68]. Thus, let $y = \mathscr{A} x$, there exists $\beta > 0$ such that $\left\| \mathscr{A}^{-1} y \right\|_{\mathbb{X}} \leqslant \beta \left\| y \right\|_{\mathbb{Y}} \implies \left\| x \right\|_{\mathbb{X}} \leqslant \beta \left\| \mathscr{A} x \right\|_{\mathbb{Y}}$, and hence $\mathscr{A}$ is bounded below. ∎

REFERENCES

[1] I. A. I. A. . ABRAMOVICH, *An invitation to operator theory*, Graduate studies in mathematics, American Mathematical Society, Providence, RI, Sept. 2002.

[2] R. A. ADAMS AND J. J. F. FOURNIER, *Sobolev Spaces*, Elsevier Science, Amsterdam, Boston, Heidelberg,..., second ed., 2003.

[3] A. O. AKUNO, L. L. RAMIREZ-RAMIREZ, C. MEHTA, T. BUI-THANH, AND J. 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), http://dx.doi.org/10.2139/ssrn.4414906.

[4] N. ANTONIĆ AND K. SIMIR BURAZIN, *Graph spaces of first-order linear partial differential operators*, Mathematical Communications, 14 (2009), pp. 135–155.

[5] T. M. APOSTOL, *Calculus*, John Wiley & Sons, Nashville, TN, 2 ed., June 1967.

[6] T. ARBOGAST AND J. L. BONA, *Methods of Applied Mathematics*, University of Texas at Austin, 2008. Lecture notes in applied mathematics.

[7] S. AXLER, *Linear Algebra Done Right*, Undergraduate Texts in Mathematics, Springer International Publishing, Cham, 3rd ed. 2015 ed.

[8] S. BANACH, *Theorie des Operations Lineaires*, AMS Chelsea Publishing, Chelsea Publishing, Providence, RI, 2 ed., Jan. 1978.

[9] D. P. BERTSEKAS, *Constrained Optimization and Lagrange Multiplier Methods*, Academic Press, New York,. London, Paris, San Diego, San Francisco, 1982.

[10] D. P. BERTSEKAS, *Nonlinear Programming*, Athena Scientific, Belmont, 1999.

[11] L. 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, https://books.google.com/books?id=tflUtbrn7NQC.

[12] L. BIEGLER, O. GHATTAS, M. HEINKENSCHLOSS, AND B. WAANDERS, *Large-Scale PDE-Constrained Optimization*, Lecture Notes in Computational Science and Engineering, Springer Berlin Heidelberg, 2012, https://books.google.com/books?id=O4YKBwAAQBAJ.

[13] 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, https://books.google.com/books?id=cvwZQzwNSGsC.

[14] F. BRAUER, C. CASTILLO-CHAVEZ, AND Z. FENG, *Mathematical Models in Epidemiology*, Springer New York, 2019, https://doi.org/10.1007/978-1-4939-9828-9, https://doi.org/10.1007/978-1-4939-9828-9.

[15] *Mathematical Epidemiology*, Springer Berlin Heidelberg, 2008, https://doi.org/10.1007/978-3-540-78911-6, https://doi.org/10.1007/978-3-540-78911-6.

[16] H. Brezis, *Functional Analysis, Sobolev Spaces and Partial Differential Equations*, Universitext, Springer New York, 2010, https://books.google.nl/books?id=GAA2XqOIIGoC.

[17] T. Bui-Thanh, L. Demkowicz, and O. Ghattas, *A unified discontinuous Petrov-Galerkin method and its analysis for Friedrichs' systems*, SIAM J. Numer. Anal., 51 (2013), pp. 1933–1958. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/DPGunifiedRevised.pdf.

[18] T. Bui-Thanh and O. Ghattas, *Analysis of the Hessian for inverse scattering problems. Part I: Inverse shape scattering of acoustic waves*, Inverse Problems, 28 (2012), p. 055001, https://doi.org/10.1088/0266-5611/28/5/055001. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactI.pdf.

[19] T. Bui-Thanh and O. Ghattas, *Analysis of the Hessian for inverse scattering problems. Part II: Inverse medium scattering of acoustic waves*, Inverse Problems, 28 (2012), p. 055002, https://doi.org/10.1088/0266-5611/28/5/055002. http://users.ices.utexas.edu/%7Etanbui/PublishedPapers/CompactII.pdf.

[20] T. Bui-Thanh and O. 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.

[21] D. Colton and R. Kress, *Integral equation methods in scattering theory*, John Wiley & Sons, 1983.

[22] D. Colton and R. Kress, *Inverse Acoustic and Electromagnetic Scattering*, Applied Mathematical Sciences, Vol. 93, Springer-Verlag, Berlin, Heidelberg, New-York, Tokyo, second ed., 1998.

[23] D. J. Daley and J. Gani, *Epidemic Modelling*, Cambridge University Press, Feb. 1984, https://doi.org/10.1017/cbo9780511608834, https://doi.org/10.1017/cbo9780511608834.

[24] S. S. Demidov, *On the history of the theory of linear differential equations*, Archive for History of Exact Sciences, 28 (1983), pp. 369–387, http://www.jstor.org/stable/41133697 (accessed 2023-06-06).

[25] 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 (2009), pp. 873–885, https://doi.org/10.1098/rsif.2009.0386, https://doi.org/10.1098/rsif.2009.0386.

[26] L. C. Dilip Ghosh Roy, *Inverse Problems and Inverse Scattering of Plane Waves*, Elsevier, 2001.

[27] S. Dobson, *Epidemic modelling - Some notes, maths, and code*, Independent Publishing Network, July 2020.

[28] C. H. Edwards, *Advanced calculus of several variables*, Dover Books on Mathematics, Dover Publications, Mineola, NY, Apr. 1995.

[29] A. Ern and J.-L. Guermond, *Theory and Practice of Finite Elements*, vol. 159 of Applied Mathematical Sciences, Spinger-Verlag, 2004.

[30] A. Ern and J.-L. Guermond, *Discontinuous Galerkin methods for Friedrichs' systems. Part I. General theory*, SIAM J. Numer. Anal., 44 (2006), pp. 753–778.

[31] A. Ern, J.-L. Guermond, and G. Caplain, *An intrinsic criterion for the bijectivity of Hilbert operators related to Friedrichs' systems*, Communications in partial differential equations, 32 (2007), pp. 317–341.

[32] M. Frank and P. Wolfe, *An algorithm for quadratic programming*, Naval Research Logistics Quarterly, 3 (1956), pp. 95–110, https://doi.org/10.1002/nav.3800030109, https://doi.org/10.1002/nav.3800030109.

[33] K. O. Friedrichs, *Symmetric positive linear differential equations*, Communications on pure and applied mathematics, XI (1958), pp. 333–418.

[34] L. Fuchs, *Zur theorie der linearen differentialgleichungen mit veränderlichen coefficienten.*, Journal für die reine und angewandte Mathematik, 66 (1866), pp. 121–160, http://eudml.org/doc/147986.

[35] K. Fukushima, *Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position*, Biological Cybernetics, 36 (1980), pp. 193–202, https://doi.org/10.1007/bf00344251, https://doi.org/10.1007/bf00344251.

[36] J. Hadamard, *Sur les problèmes aux dérivés partielles et leur signification physique*, Princeton University Bulletin, 13 (1902), pp. 49–52.

[37] B. C. Hall, *Lie Groups, Lie Algebras, and Representations*, Springer International Publishing, 2015, https://doi.org/10.1007/978-3-319-13467-3, https://doi.org/10.1007/978-3-319-13467-3.

[38] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, in 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, June 2016, https://doi.org/10.1109/cvpr.2016.90, https://doi.org/10.1109/cvpr.2016.90.

[39] J. Heffernan, R. Smith, and L. Wahl, *Perspectives on the basic reproductive ratio*, Journal of The

Royal Society Interface, 2 (2005), pp. 281–293, https://doi.org/10.1098/rsif.2005.0042, https://doi.org/10.1098/rsif.2005.0042.

[40] H. W. HETHCOTE, *The mathematics of infectious diseases*, SIAM Review, 42 (2000), pp. 599–653, https://doi.org/10.1137/s0036144500371907, https://doi.org/10.1137/s0036144500371907.

[41] R. A. HORN AND C. A. JOHNSON, *Topics in Matrix Analysis*, Cambridge University Press, Cambridge, London, New York, 1991.

[42] D. H. HUBEL AND T. N. WIESEL, *Receptive fields and functional architecture of monkey striate cortex*, The Journal of Physiology, 195 (1968), pp. 215–243, https://doi.org/10.1113/jphysiol.1968.sp008455, https://doi.org/10.1113/jphysiol.1968.sp008455.

[43] A. HURFORD, D. COWNDEN, AND T. DAY, *Next-generation tools for evolutionary invasion analyses*, Journal of The Royal Society Interface, 7 (2009), pp. 561–571, https://doi.org/10.1098/rsif.2009.0448, https://doi.org/10.1098/rsif.2009.0448.

[44] A. J. AND C. J., *Generalization of the Wolfe Reduced Gradient Method to the Case of Nonlinear Constraints in Optimization*, Academic Press, 1969.

[45] M. JENSEN, *Discontinuous Galerkin methods for Friedrichs' systems with irregular solutions*, PhD thesis, University of Oxford, 2004.

[46] J. KIEFER AND J. WOLFOWITZ, *Stochastic estimation of the maximum of a regression function*, The Annals of Mathematical Statistics, 23 (1952), pp. 462–466, https://doi.org/10.1214/aoms/1177729392, https://doi.org/10.1214/aoms/1177729392.

[47] M. KLINE, *Mathematical thought from ancient to modern times*, Oxford University Press, New York, NY, Sept. 1972.

[48] J. L. LAGRANGE, *Solution de différents problèmes de Calcul intégral*, M. A. T. 3, 3 (1766), pp. 179–380.

[49] S. LANG, *Fundamentals of differential geometry*, Graduate Texts in Mathematics, Springer, New York, NY, 1 ed., Sept. 2001.

[50] Y. LECUN, Y. BENGIO, AND G. HINTON, *Deep learning*, Nature, 521 (2015), pp. 436–444, https://doi.org/10.1038/nature14539, https://doi.org/10.1038/nature14539.

[51] G. LEDDER, *Mathematical Modeling for Epidemiology and Ecology*, Springer International Publishing, 2023, https://doi.org/10.1007/978-3-031-09454-5, https://doi.org/10.1007/978-3-031-09454-5.

[52] *Constrained Optimization and Optimal Control for Partial Differential Equations*, Springer Basel, Oct. 2011, https://doi.org/10.1007/978-3-0348-0133-1, https://doi.org/10.1007/978-3-0348-0133-1.

[53] J. LINDSTRÖM, *On the origin and early history of functional analysis*, 2008.

[54] S. 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, (1970), pp. 6–7.

[55] S. LINNAINMAA, *Taylor expansion of the accumulated rounding error*, BIT, 16 (1976), pp. 146–160, https://doi.org/10.1007/bf01931367, https://doi.org/10.1007/bf01931367.

[56] H. LOGEMANN AND E. P. RYAN, *Ordinary Differential Equations Analysis, Qualitative Theory and Control*, Springer undergraduate mathematics series, Springer, London, 2014.

[57] J. C. D. LOS REYES, *Numerical PDE-Constrained Optimization*, Springer International Publishing, 2015, https://doi.org/10.1007/978-3-319-13395-9, https://doi.org/10.1007/978-3-319-13395-9.

[58] D. G. LUENBERGER, *Optimization by Vector Space Methods*, John Wiley and Sons, New York, 1969.

[59] M. MARTCHEVA, *An Introduction to Mathematical Epidemiology*, Springer US, 2015, https://doi.org/10.1007/978-1-4899-7612-3, https://doi.org/10.1007/978-1-4899-7612-3.

[60] W. McLEAN, *Strongly Elliptic Systems and Boundary Integral Equations*, Cambidge University Press, 2000.

[61] C. MOLER AND C. V. LOAN, *Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later*, SIAM Review, 45 (2003), pp. 3–49, https://doi.org/10.1137/s00361445024180, https://doi.org/10.1137/s00361445024180.

[62] J. NOCEDAL AND S. J. WRIGHT, *Numerical Optimization*, Springer Verlag, Berlin, Heidelberg, New York, second ed., 2006.

[63] J. T. ODEN AND L. F. DEMKOWICZ, *Applied functional analysis*, CRC Press, 2010.

[64] A. PERASSO, *An introduction to the basic reproduction number in mathematical epidemiology*, ESAIM: Proceedings and Surveys, 62 (2018), pp. 123–138, https://doi.org/10.1051/proc/201862123, https://doi.org/10.1051/proc/201862123.

[65] F. RIESZ, *Über lineare funktionalgleichungen*, Acta math, 41 (1918), p. 8.

[66] H. Robbins and S. Monro, *A stochastic approximation method*, The Annals of Mathematical Statistics, 22 (1951), pp. 400–407, https://doi.org/10.1214/aoms/1177729586, https://doi.org/10.1214/aoms/1177729586.

[67] H. Royden and P. Fitzpatrick, *Real Analysis (Classic Version)*, Pearson, Upper Saddle River, NJ, 4 ed., Feb. 2017.

[68] W. Rudin, *Functional Analysis*, McGraw-Hill, New York, St. Louis, San Francisco,..., 1973.

[69] J. Schmidhuber, *Deep learning in neural networks: An overview*, Neural Networks, 61 (2015), pp. 85–117, https://doi.org/10.1016/j.neunet.2014.09.003, https://doi.org/10.1016/j.neunet.2014.09.003.

[70] A. Shapiro, D. Dentcheva, and A. Ruszczynski, *Lectures on Stochastic Programming: Modeling and Theory*, Society for Industrial and Applied Mathematics, 2009.

[71] R. Showalter, *Hilbert Space Methods in Partial Differential Equations*, vol. Monograph, 06 1998.

[72] R. E. Showalter, *Hilbert Space Methods for Partial Differential Equations*, Pitman, London, San Francisco, Melbourne, 1977.

[73] G. Strang, *The fundamental theorem of linear algebra*, Am. Math. Mon., 100 (1993), p. 848.

[74] A. N. Tikhonov and V. Y. Arsenin, *Solutions of ill-posed problems*, Scripta series in mathematics, John Wiley & Sons, Nashville, TN, Aug. 1977.

[75] A. 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 ed., June 1995.

[76] P. van den Driessche and J. Watmough, *Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission*, Mathematical Biosciences, 180 (2002), pp. 29–48, https://doi.org/10.1016/s0025-5564(02)00108-6, https://doi.org/10.1016/s0025-5564(02)00108-6.

[77] K. Yosida, *Functional Analysis*, Springer Berlin Heidelberg, 1995, https://doi.org/10.1007/978-3-642-61859-8, https://doi.org/10.1007/978-3-642-61859-8.