Appendix A

Banach spaces and Hilbert spaces

This appendix consists of definitions and essential theorems on Banach and Hilbert spaces and norms. The organization loosely follows the chapters on Banach and Hilbert spaces in Renardy and Rogers [396], and we recommend that text to the reader for a more thorough treatment and for the proofs we have omitted for brevity. We assume some background in analysis at the level of Rudin and familiarity with the L^p spaces as well.

Definition A.0.1. *A normed linear space that is complete in its metric is called a* Banach space. *That is, X is a Banach space if there exists a sequence $\{x_n\} \subset X$ with $\|x_n - x_m\| \to 0$ as $m,n \to \infty$; then $\{x_n\}$ converges in X.*

Recall that such a sequence is called a *Cauchy sequence*, and so a Banach space is a normed linear space in which all Cauchy sequences converge to an element in the space.

Examples of Banach spaces include the following:

- Bounded continuous functions on the closure of an open set $\Omega \subset \mathbb{R}^m$, denoted by $C_b(\bar{\Omega})$, equipped with norm $\|u\| = \sup_{x\in\bar{\Omega}} |u(x)|$.
- Bounded analytic functions on an open set $\Omega \subset \mathbb{C}$ with norm $\|u\| = \sup_{x\in\Omega} |u(z)|$.
- The L^p spaces, $1 \le p < \infty$.

Definition A.0.2. *Let X be a vector space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. An* inner product *on X is a function $\langle\cdot,\cdot\rangle : X \times X \to \mathbb{K}$ that, for all $x,y,z \in X$ and for all $\alpha \in \mathbb{K}$, satisfies*

1. $\langle x,y\rangle = \overline{\langle y,x\rangle}$,
2. $\langle \alpha x,y\rangle = \alpha\langle x,y\rangle$,
3. $\langle x+y,z\rangle = \langle x,z\rangle + \langle y,z\rangle$,
4. $\langle x,x\rangle \ge 0$, *and* $\langle x,x\rangle = 0$ *if and only if* $x = 0$.

The space $(X,\langle\cdot,\cdot\rangle)$ is called an inner product space.

Definition A.0.3. *An inner product space that is complete in its metric is called a* Hilbert space.

Definition A.0.4. *Two norms* $\|\cdot\|_1$ *and* $\|\cdot\|_2$ *are* equivalent *if there exist constants* $m, M > 0$ *such that*

$$m\|x\|_1 \leq \|x\|_2 \leq M\|x\|_1 \qquad \textit{for all} \ \ x \in X.$$

Definition A.0.5. *Elements* x, y *of a Hilbert space* H *are* orthogonal *if* $\langle x, y\rangle = 0$.

Definition A.0.6. *Let* M *be a linear subspace of a Hilbert space* H. *We define the* orthogonal complement of M *by* $M^{\perp}$ *by*

$$M^{\perp} = \{x \in H : \langle x, y\rangle = 0 \ \ \textit{for all} \ \ y \in M\}.$$

Note that the orthogonal complement is always a closed subspace.

Theorem A.1 (projection theorem). *Let* M *be a closed linear subspace of a Hilbert space* H. *Then every* $x \in H$ *has the unique decomposition* $x = y + z$ *where* $y \in M$ *and* $z \in M^{\perp}$.

Appendix B

Mappings and compact operators

Theorem B.1. *Let L be a linear mapping from a normed linear space X to a normed linear space Y. Then the following are equivalent:*

1. *L is continuous at* 0.
2. *L is continuous.*
3. *There exists $c \geq 0$ such that $\|Lx\| \leq c\|x\|$ for all $x \in X$ (boundedness).*

Item 3 above is the definition of a bounded linear mapping. If L is a bounded linear mapping from X to Y, we define its norm by

$$\|L\| = \sup_{\|x\|=1} \|Lx\|. \tag{B.1}$$

Definition B.0.7. *Let $\mathcal{L}(X,Y)$ denote the space of bounded linear mappings from X to Y.*

Note that if X and Y are Banach spaces, then $\mathcal{L}(X,Y)$ equipped with the norm (B.1) is also a Banach space.

Theorem B.2 (Riesz representation theorem). *Let $L : X \to \mathbb{C}$ be a bounded linear functional. Then there exists a unique $x_0 \in X$ such that $Lx = \langle x, x_0 \rangle$ for all $x \in X$. Moreover,*

$$\|L\| = \|x_0\|.$$

Definition B.0.8. *A* linear operator *from a Banach space X to a Banach space Y is the pair $(D(A), A)$, where $D(A) \subset X$ is called the domain of A and A is a linear transformation from $D(A)$ to Y. The* range *of A is the subspace*

$$\text{range}(A) = \{y \in Y : y = Ax \ \textit{ for some } \ x \in D(A)\}.$$

The nullspace *of A is the subspace*

$$\mathcal{N}(A) = \{x \in X : Ax = 0\}.$$

Theorem B.3. *Let A be a linear operator from a Banach space X to a Banach space Y with domain $D(A)$ and range* range(A)*. Then the following two properties hold:*

1. *The inverse operator A^{-1} : range$(A) \to X$ exists if and only if $\mathcal{N}(A) = \{0\}$.*

2. *If the inverse operator exists, it is linear.*

Compact operators play an important role in inverse problems. We include several essential definitions and theorems. Most proofs are omitted for brevity, and the reader is again referred to [396] for a more complete exposition. The following theorem provides an important characterization of compact operators.

Theorem B.4. *The operator $K : D(K) \subset X \to Y$ is compact if and only if for every bounded sequence $\{x_n\}$ in $D(K)$, $\{Kx_n\}$ has a convergent subsequence.*

Theorem B.5. *Every compact operator is bounded.*

Proof. Assume K is an unbounded operator. Then there exists $\{u_n\} \subset D(K)$ with $\|u_n\| = 1$ such that $\lim_{n\to\infty} \|Ku\| = \infty$. Choose elements of $\{u_n\}$ to create a subsequence $\{\tilde{u}_n\}$ such that $\{K\tilde{u}_n\}$ is (strictly) monotone and $\lim_{n\to\infty} \|K\tilde{u}_n\| = \infty$. Note that since K is compact and $\|\tilde{u}_n\| = 1$, $\{K\tilde{u}_n\}$ must contain a convergent subsequence. This is a contradiction and proves that K is bounded. □

Note: Let H be a Hilbert space. If $K : D(K) \subset H \to H$ is compact, then K is bounded on $D(K)$ and there exists an extension $\tilde{K} \subset \mathcal{L}(H)$ such that $||\tilde{K}\| = \|K\|$ and $\tilde{K} = K$ on $D(K)$. It can be shown that $\tilde{K}$ is compact. Thus if K is compact, we can take $D(K) = H$.

Note: If $K \in \mathcal{L}(H,H)$, K is not necessarily compact. As an example, consider the identity operator I on an infinite-dimensional Hilbert space H. The identity operator I is not compact.

Proof. Suppose I is compact. Let $\{e_n\}$ be an orthonormal basis of H. Since $\|e_n\| = 1$, $\{Ie_n\}$ must have a convergence subsequence $\{e'_n\}$. But this is impossible since $\|e'_n - e_n\|^2 = \langle e'_n - e_n, e'_n - e_n\rangle = \|e'_n\|^2 + \|e_n\|^2 - 2\Re\langle e'_n, e_n\rangle = 2$. Thus, no subsequence can be Cauchy. So I is not compact. □

Theorem B.6. *Let $A : D(A) \subset X \to Y$ be a linear operator.*

1. *If A is bounded and* range(A) *is finite-dimensional, then A is compact.*

2. *If $D(A)$ is finite-dimensional, then A is compact.*

***P*roof.**

1. Let $\{x_n\} \subset D(A)$ with $\|x_n\| \leq M$ for all n. Then $\{Ax_n\} \subset$ range(A) and $\|Ax_n\| \leq \|A\|M \leq \infty$. Since range$(A)$ is a finite-dimensional Hilbert space, by the Bolzano–Weierstrass theorem, $\{Ax_n\}$ has a convergent subsequence. So A is compact.

2. Let $\{e_n\}_{n=1}^N$ be an orthonormal basis for $D(A)$, $(N < \infty)$. Then

$$\text{linear span}\{Ae_n\}_{n=1}^N = \text{range}(A).$$

Thus, $\dim \mathcal{R}(A) \leq \dim D(A)$. So $A : D(A) \to \text{range}(A)$ which has dimension $N' \leq N$. Thus, A is a bounded operator. So A is compact. $\square$

Recall Bessel's inequality: If $\{e_n\}$ is a orthonormal set in H, then for all $u \in H$, $\sum_{n=1}^{\infty} |\langle u, e_n\rangle|^2 \leq \|u\|^2$. Thus, whenever $\{e_n\}$ is an orthonormal set in H, then $\lim_{n\to\infty}\langle u, e_n\rangle = 0$ for all $u \in H$ since the above sum converges.

Definition B.0.9. *A sequence $\{f_n\}$ is said to converge weakly to $f \in H$ if $\langle f_n, u\rangle \to \langle f, u\rangle$ as $n \to \infty$ for all $u \in H$. We write this as $f_n \overset{wk}{\to} f$.*

Theorem B.7. *If $f_n \to f$ in H (i.e., $\|f_n - f\| \to 0$), then $(f_n \overset{wk}{\to} f)$ in H.*

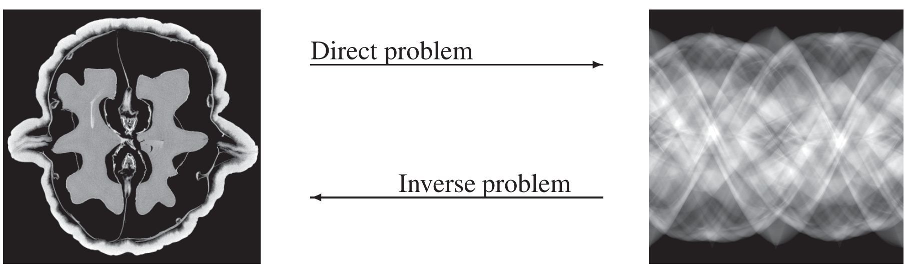
Proof. Let $u \in H$. Then $|\langle f_n, u\rangle - \langle f, u\rangle| = |\langle f_n - f, u\rangle| \leq \|f_n - f\|\|u\| \to 0$ as $n \to \infty$. This completes the proof. $\square$

Chapter 1

Introduction

Inverse problems are the opposites of direct problems. Informally, in a direct problem one finds an effect from a cause, and in an inverse problem one is given the effect and wants to recover the cause. The most usual situation giving rise to an inverse problem is the need to interpret indirect physical measurements of an unknown object of interest.

For example, in X-ray tomography the direct problem is to determine the X-ray projection images we would get from a physical body whose internal structure we know precisely. The corresponding inverse problem is to reconstruct the inner structure of an unknown physical body from the knowledge of X-ray images taken from different directions. Here is a two-dimensional example:

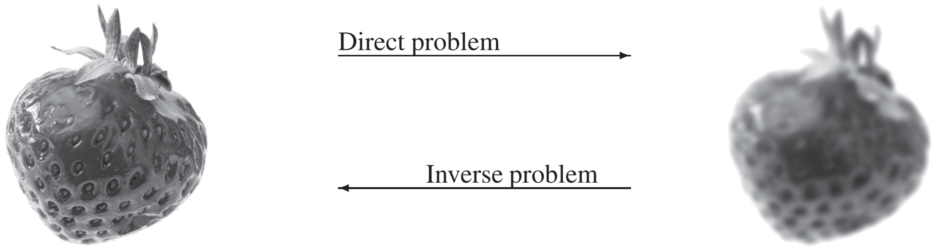


Here the slice through a walnut (left) is the cause and the collection of X-ray data (right) is the effect. The tomographic data is shown in the traditional *sinogram* form, which will be discussed in detail in Section 2.3.5. The slice image on the left is courtesy of Keijo Hämäläinen and Aki Kallonen from University of Helsinki, Finland.

Variants of the above tomographic problem appear also in monitoring ozone profiles in the upper atmosphere using spaceborne star occultation measurements [296], identifying molecules based on electron microscope imaging [139], interpretation of Doppler weather radar measurements [299], and measuring the temperature distribution of hot gases flowing through the window of a burning house using metal wires [30]. This demonstrates the general nature of mathematics: the same underlying problem may be found in very different application areas.

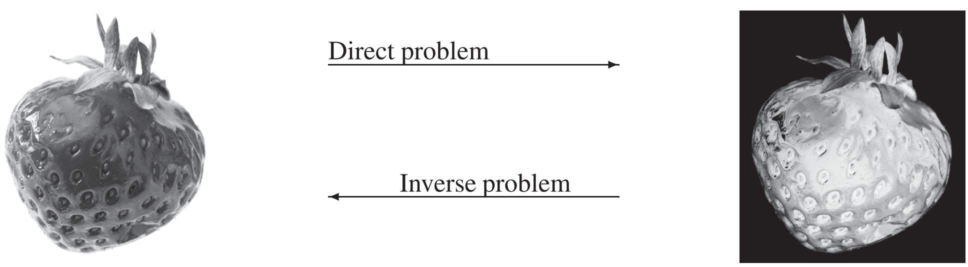
Another example comes from image processing. Define the direct problem as finding out how a given sharp photograph would look like if the camera were incorrectly focused.

The inverse problem known as *deblurring* is finding the sharp photograph from a given blurry image.



Here the cause is the sharp image and the effect is the blurred image. A famous example of deblurring is the correction algorithm used for the Hubble Space Telescope images after finding out a flaw in the construction of its main lens.

There is an apparent symmetry in the above explanation: without further restriction of the definitions, the direct problem and inverse problem would be in identical relation with each other. For example, we might take as the direct problem the determination of a positive photograph from the knowledge of the negative photograph.



In this case the corresponding "inverse problem" would be inverting a given photograph to arrive at the negative. Here both problems are easy and stable, and one can move between them repeatedly.

However, we concentrate on *ill-posed inverse problems*, where the inverse problem is more difficult to solve than the direct problem. To explain this we need the notion of a *well-posed problem* introduced by Jacques Hadamard (1865–1963):

H_1: **Existence.** There should be at least one solution.

H_2: **Uniqueness.** There should be at most one solution.

H_3: **Stability.** The solution must depend continuously on data.

Now denote by $\mathcal{A}$ the *forward map*, defined conceptually by $\mathcal{A}(\text{cause}) = \text{effect}$. The direct problem must be well-posed, in other words $\mathcal{A}$ should be a well-defined, single-valued, and continuous function. The inverse problem is ill-posed if $\mathcal{A}^{-1}$ does not exist or is not continuous; then at least one of the conditions H_1–H_3 fails for $\mathcal{A}^{-1}$. In the positive-negative photograph example above both $\mathcal{A}$ and $\mathcal{A}^{-1}$ are well-defined and continuous, so it is not an ill-posed inverse problem.

Our general mathematical model is constructed as follows. We consider indirect linear measurements of the form

$$\mathbf{m} = \mathcal{A} f + \varepsilon, \tag{1.1}$$

where f is a piecewise continuous function defined on a subset of $\mathbb{R}^d$ and $\mathbf{m} \in \mathbb{R}^k$ is a vector of numbers given by a measurement device. Models of the form (1.1) arise from various situations in technology or physics; the linear operator $\mathcal{A}$ may be related, for example, to a PDE or to an integral equation. We will discuss examples of practically relevant operators $\mathcal{A}$ in Chapter 2.

The vector $\varepsilon \in \mathbb{R}^k$ in (1.1) models errors coming from measurement noise, which is inevitable in practical situations. Sometimes ε is modeled as a random variable with certain statistics. However, in this book we think of ε as a deterministic but unknown error; the information we have on ε is an inequality $\|\varepsilon\| \leq \delta$ with a known constant $\delta > 0$. Such a number δ can often be found by calibration of the measurement device; higher-quality device typically gives smaller δ. Our deterministic approach does include the possibility that ε is a fixed realization of a random process; this is actually a quite accurate model of many realistic measurements.

The reason for the term "indirect measurement" is the following. We are interested in the function f but cannot measure its values directly. However, f is connected to another physical quantity $\mathbf{m}$, which is available for measurement. The connection is modeled by the linear operator $\mathcal{A}$. Now the direct problem is "Given f, determine $\mathbf{m} = \mathcal{A} f$." The corresponding inverse problem is

$$\begin{array}{l}\text{Given noisy measurement } \mathbf{m} = \mathcal{A} f + \varepsilon \text{ and } \delta > 0 \\ \text{with } \|\varepsilon\| \leq \delta, \text{ extract information about } f.\end{array} \tag{1.2}$$

Part I of this book is about practical extraction of information from indirect linear measurements using computational methods. Consequently, we need to introduce a finite-dimensional approximation $\mathbf{f} \in \mathbb{R}^n$ to the function f and to build a matrix model for the linear operator $\mathcal{A}$.

In case of discrete linear inverse problems we consider measurements of the form

$$\mathbf{m} = A\mathbf{f} + \varepsilon, \tag{1.3}$$

where $\mathbf{m} \in \mathbb{R}^k$ and $\mathbf{f} \in \mathbb{R}^n$. Moreover, A is a matrix of size $k \times n$ (k rows and n columns). Strictly speaking, we abuse notation by using $\mathbf{m}$ in both (1.1) and (1.3) although they are different models. Whenever there is the possibility of confusion, we denote the measurement from the finite model (1.3) by $\mathbf{m}^{(n)}$.

Once the computational model (1.3) has been constructed, it is tempting to try to solve the inverse problem (1.2) by the naïve reconstruction

$$\mathbf{f} \approx A^{-1}\mathbf{m}. \tag{1.4}$$

However, in the case of ill-posed inverse problems the approach (1.4) will fail. In Sections 2.1–2.3 we describe some important indirect measurements and demonstrate the failure of the naïve reconstruction (1.4) numerically.

Regularization is what really needs to be done for successful and noise-robust solution of linear inverse problems. We discuss the theory and implementation of various regularization methods in Chapters 3–7. We demonstrate the properties of the various methods using the example problems developed in Chapter 2. Large-scale computational methods are emphasized throughout the text because practical applications often lead to very high-dimensional problems.

Discretization of the continuum model (1.1) using discrete models of the form (1.3) involves choosing the dimension n of the discrete vector $\mathbf{f}$. It is desirable to design computational inversion methods that give consistent results at different resolutions n. This so-called *discretization invariance* is discussed in Chapter 8.

Practical examples of linear inversion are described in Chapter 9 in the case of X-ray tomography.

It is impossible to cover all useful and important material related to computational inversion in this book. We list here some further reading that complements our approach.

Regularization theory is discussed in the classical texts [439, 340] and in the more recent books [132, 254, 468]. The mathematical foundations of inverse problems are explained more generally in the books [236, 269].

Some of the computational inversion methods presented in this book (such as truncated singular value decomposition, Tikhonov, regularization, and total variation regularization) are discussed also in [461, 194, 196, 195, 354]. Useful methodologies that are not covered here due to restrictions of space include truncated iterative solvers [193, 191, 65], approximate inverses [407], statistical inversion [247, 434, 67], and variational methods [370, 405].

Application-oriented texts are available for inverse problems in medical imaging [249, 353, 133], geophysical inversion [331, 434], and signal processing [403].