Matt Dunlop, James Robinson & Andrew Stuart

Introduction to Linear Analysis with Applications

CMS/ACM/IDS 107 Lecture Notes

Contents

Preface

These lecture notes give an introduction to the subject of linear analysis. Although the treatment is a rigorous mathematical one, the aim is to overview the basic material in a way that is useful for applications. In particular the aim is to lay the foundations needed to enable formulation of important problems in science and engineering, to understand the language used in these formulations, and to understand key concepts and their inter-relations. Knowing how proofs work is intrinsically appealing, but also plays an important role in the understanding of new ideas. Thus, although we do not prove everything that we state, the majority of results do come equipped with full proofs.

Some of the material in these notes is modified from Warwick University lecture notes on "Matrix Analysis and Algorithms", co-authored by Andrew Stuart, Tim Sullivan and Jochen Voss, and on "Linear Partial Differential Equations", by James Robinson. Further references, both classic and modern, include Griffel (2002), Hutson et al. (2005), Luenberger (1969), Robinson (2020) and Zeidler (1995). Part of these notes concerns linear functional analysis related to functions defined on subsets of $\mathbb{R}^d$; this will be touched on in the books just cited, but texts concerning differential equations will contain much more on the subject. Parts of the comprehensive text Evans (2002) will be useful in this context, and the text Adams and Fournier (2003) has a deep presentation of some of the function spaces studied here. The text Hanke (2017) offers an inverse problems flavor to the subject.

The authors are also grateful to the students of CMS/ACM/IDS 107 in academic years 16/17, 17/18, 18/19, 19/20 and 20/21 for their input, which helped improve the notes, and to Dmitry Burov, Nikola Kovachki and Nicholas Nelsen, who acted as TAs and also helped improve the notes.

1
Data Science Motivation

This lecture demonstrates the role of linear analysis in the formulation of a basic task from data science – clustering of data. We concentrate on methods which exploit the graph Laplacian.

1.1 Problem Statement

We are given data in the form of points $\{x^{(n)}\}_{n\in \mathsf{Z}}$ in a metric space (X, d), $\mathsf{Z} \coloneqq \{1, \dots, N\}$. Our aim is to assign a label $y^{(n)} \in \{\pm 1\}$ to each $n \in \mathsf{Z}$ by using correlation in the data; that is, the aim is to cluster the points into two distinct classes defined by label ± 1. The ideas presented in this lecture can be extended to the more general problem considering $k \geq 2$ classes. Furthermore there are many ways one can approach this problem including Principal Component Analysis (PCA), k-mean clustering, and, as presented here, graph-based spectral clustering.

To be concrete you may think of the setting where $X = \mathbb{R}^q$ and the metric $d(x, x') = \|x - x'\|$, where $\|\cdot\|$ is the Euclidean norm.

1.2 Graph Laplacian

View Z as the nodes (vertices) of an undirected graph. Let

$$\mathsf{E} = \{(i, j) \colon (i, j) \in \mathsf{Z} \times \mathsf{Z}\}$$

be the corresponding edge set. Note that a function $\phi \colon \mathsf{Z} \to \mathbb{R}$ can be viewed as a vector $\phi \in \mathbb{R}^N$ and similarly a function $W : \mathsf{E} \to \mathbb{R}$ can be viewed as a matrix $W \in \mathbb{R}^{N\times N}$. We will choose a W to measure the affinity between nodes, corresponding to a weighted adjacency matrix.

Definition 1.1. *The weight function $\eta : \mathbb{R}^+ \to \mathbb{R}^+$ satisfies:*

(i) $\lim_{r\to\infty} \eta(r) = 0$;
(ii) $\eta(r_1) \le \eta(r_2)$ *if* $r_1 \ge r_2$.

Remark 1.2. *The weight function is often normalized by imposing conditions such as:*

(i) $\eta(0) = 1$,
(ii) $\int_{\mathbb{R}^+} \eta(r)dr = 1$.

Definition 1.3. *Given a metric d on X, define the weighted adjacency matrix $W \in \mathbb{R}^{N\times N}$ by*

$$W_{ij} = \eta\big(d(x^{(i)}, x^{(j)})\big).$$

We note that W is symmetric because any metric $d(\cdot,\cdot)$ is symmetric.

Definition 1.4. *The degree matrix $D \in \mathbb{R}^{N\times N}$ is defined by*

$$D_{ij} = \delta_{ij} \sum_{k\in \mathsf{Z}} W_{ik},$$

where δ_{ij} is Kronecker delta. In particular, $D_{ii} = \sum_{k\in\mathsf{Z}} W_{ik}$ and $D_{ij} = 0$ if $i \ne j$.

Definition 1.5. *The (unnormalized) graph Laplacian $L \in \mathbb{R}^{N\times N}$ is defined as the symmetric matrix*

$$L = D - W.$$

Theorem 1.6. *The graph Laplacian L satisfies:*

(i) *L is positive semi-definite:*

$$\langle \phi, L\phi\rangle = \frac{1}{2} \sum_{(i,j)\in \mathsf{E}} W_{ij}|\phi_i - \phi_j|^2 \ge 0, \quad \forall \phi \in \mathbb{R}^N.$$

(ii) *If $\mathbb{1} = (1, 1, \dots, 1)^\top \in \mathbb{R}^N$ then $L\mathbb{1} = 0$.*
(iii) *If $W_{ij} > 0 \ \ \forall (i,j) \in \mathsf{E}$ (we then say that the graph is* fully connected*) then $\mathbb{1}$ is the only eigenvector with eigenvalue zero.*

Proof For the first property,

$$\begin{aligned}
\langle \phi, L\phi \rangle &= \sum_{(i,j)\in \mathsf{E}} \phi_i L_{ij} \phi_j \\
&= \sum_{(i,j)\in \mathsf{E}} \phi_i D_{ij} \phi_j - \sum_{(i,j)\in \mathsf{E}} \phi_i W_{ij} \phi_j \\
&= \sum_{(i,k)\in \mathsf{E}} \phi_i^2 W_{ik} - \sum_{(i,j)\in \mathsf{E}} \phi_i W_{ij} \phi_j \\
&= \frac{1}{2} \sum_{(i,k)\in \mathsf{E}} \phi_i^2 W_{ik} + \frac{1}{2} \sum_{(j,k)\in \mathsf{E}} \phi_j^2 W_{jk} - \sum_{(i,j)\in \mathsf{E}} \phi_i W_{ij} \phi_j \\
&= \frac{1}{2} \sum_{(i,j)\in \mathsf{E}} \phi_i^2 W_{ij} + \frac{1}{2} \sum_{(i,j)\in \mathsf{E}} \phi_j^2 W_{ji} - \sum_{(i,j)\in \mathsf{E}} \phi_i W_{ij} \phi_j \\
&= \frac{1}{2} \sum_{(i,j)\in \mathsf{E}} \phi_i^2 W_{ij} + \frac{1}{2} \sum_{(i,j)\in \mathsf{E}} \phi_j^2 W_{ij} - \sum_{(i,j)\in \mathsf{E}} \phi_i W_{ij} \phi_j \\
&= \frac{1}{2} \sum_{(i,j)\in \mathsf{E}} W_{ij} |\phi_i - \phi_j|^2
\end{aligned}$$

as desired. For the second property,

$$\begin{aligned}
(L\mathbb{1})_i &= \sum_{j=1}^{N} L_{ij} \\
&= \sum_{j=1}^{N} D_{ij} - \sum_{j=1}^{N} W_{ij} \\
&= D_{ii} - \sum_{j=1}^{N} W_{ij} \\
&= D_{ii} - D_{ii} \\
&= 0
\end{aligned}$$

as desired. For the third property, first note that all eigenvectors are real since L is symmetric. Suppose, for contradiction, that there is an eigenvector ϕ with corresponding eigenvalue zero which is not proportional to $\mathbb{1}$. Then $\exists (k, \ell) \in \mathsf{E}$ such that $\phi_k \neq \phi_\ell$. If ϕ were an eigenvector corresponding to eigenvalue zero then $L\phi = 0$. But then, by property one and the symmetry of W,

$$0 = \langle \phi, 0 \rangle = \langle \phi, L\phi \rangle \geq W_{k\ell} |\phi_k - \phi_\ell|^2 > 0$$

a contradiction. Thus ϕ is not an eigenvector with eigenvalue zero. □

Definition 1.7. *The normalized graph Laplacian $L' \in \mathbb{R}^{N\times N}$ is defined by*

$$L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}.$$

Remark 1.8. *Note that $D_{ii} \geq W_{ii} = \eta(0) = 1$ hence we may define D^α for any $\alpha \in \mathbb{R}$ by $(D^\alpha)_{ii} = (D_{ii})^\alpha$.*

Theorem 1.9. *L' is symmetric positive semi-definite and has a zero eigenvalue with eigenvector $D^{\frac{1}{2}}\mathbb{1}$.*

Proof Left as Exercise 1.5. □

1.3 Clustering by Graph Laplacian

We first consider an idealized case.

Assumption 1.10. *We make the following assumptions:*

(i) $\mathsf{Z} = \mathsf{Z}^+ \cup \mathsf{Z}^-$ *with* $\mathsf{Z}^+ \cap \mathsf{Z}^- = \emptyset$, *and* $|\mathsf{Z}^\pm| = N^\pm$.
(ii) $W_{ij} = 0$ *for* $(i,j) \in \mathsf{Z}^+ \times \mathsf{Z}^-$.
(iii) *Graph Laplacians $L^\pm$ on $\mathsf{Z}^\pm$ are fully connected.*

Theorem 1.11. *Under Assumptions 1.10, the graph Laplacian L_0 on Z satisfies:*

(i) $L_0 = \begin{bmatrix} L^+ & 0 \\ 0 & L^- \end{bmatrix}$.
(ii) $L_0\mathbb{1} = 0$.
(iii) *If $\varphi = (\mathbb{1}^\top, -\mathbb{1}^\top)^\top$, then $L_0\varphi = 0$.*
(iv) *$\mathbb{1}, \varphi$ and linear combinations are the only eigenvectors of L_0 with eigenvalue zero.*

Proof Part (i) follows from the fact that $W_{ij} = 0$ for $(i,j) \in \mathsf{Z}^+ \times \mathsf{Z}^-$ by assumption, whilst $D_{ij} = 0$ for $(i,j) \in \mathsf{Z}^+ \times \mathsf{Z}^-$since D is diagonal by construction. Part (ii) follows directly from Theorem 1.6. The same theorem also implies that $L^\pm\mathbb{1} = 0$ since $L^\pm$ are graph Laplacians. Then for part (iii) we note that

$$\begin{bmatrix} L^+ & 0 \\ 0 & L^- \end{bmatrix} \begin{bmatrix} \mathbb{1} \\ -\mathbb{1} \end{bmatrix} = \begin{bmatrix} L^+\mathbb{1} \\ -L^-\mathbb{1} \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}.$$

For part (iv),

$$\langle \phi, L_0\phi \rangle = \langle \phi^+, L^+\phi^+ \rangle + \langle \phi^-, L^-\phi^- \rangle$$

if $\phi = ((\phi^+)^\top, (\phi^-)^\top)^\top$. Thus $\langle \phi, L_0\phi\rangle = 0$ if and only if $\phi^+ \propto \mathbb{1}$ and $\phi^- \propto \mathbb{1}$ since $L^\pm$ are fully connected and we may apply Theorem 1.6. Thus

$$\phi = \begin{bmatrix} a\mathbb{1} \\ b\mathbb{1} \end{bmatrix}$$

for some $a, b \in \mathbb{R}$. Hence $\phi \in \text{span}\{\mathbb{1}, \varphi\}$. □

We now move to a perturbed setting.

Assumption 1.12. *We make the following assumptions about a graph Laplacian L which is a small $0 < \varepsilon \ll 1$ perturbation of L_0:*

(i) $L = L_0 + \varepsilon L_1$, *with L_1 symmetric and L_0 defined with $L^\pm$ fully connected.*
(ii) $L_1\mathbb{1} = 0$.
(iii) $\langle \phi, L_1\phi\rangle > 0 \ \ \forall \phi \perp \mathbb{1}$.

Theorem 1.13. *Let Assumptions 1.12 hold and consider the eigenvalue problem*

$$L\phi^{(j)} = \lambda^{(j)}\phi^{(j)}, \quad j \in \mathsf{Z}$$

with $\{\phi^{(i)}\}_{i\in\mathsf{Z}}$ an orthogonal set and $0 = \lambda^{(1)} \le \lambda^{(2)} \le \cdots \le \lambda^{(N)}$. It follows that

(i) $\phi^{(1)} = \mathbb{1}$.
(ii) $(\phi^{(2)}, \lambda^{(2)})$ *admit the expansions*

$$\phi^{(2)} = \varphi_0 + \varepsilon\varphi_1 + O(\varepsilon^2)$$
$$\lambda^{(2)} = \varepsilon\lambda_1 + O(\varepsilon^2)$$

where $\varphi_0 = (a\mathbb{1}^\top, b\mathbb{1}^\top)^\top$, $\lambda_1 = \langle \varphi_0, L_1\varphi_0\rangle$ and a, b are the unique (up to sign) solution of

$$a^2N^+ + b^2N^- = 1$$
$$aN^+ + bN^- = 0.$$

Proof Let $(\phi^{(2)}, \lambda^{(2)}) \mapsto (\phi, \lambda)$ to simplify notation. Then

$$(L_0 + \varepsilon L_1)\phi = \lambda\phi$$

and $\langle \mathbb{1}, \phi\rangle = 0$. Enforce $\|\phi\|^2 = 1$ and set

$$\phi = \varphi_0 + \varepsilon\varphi_1 + O(\varepsilon^2)$$
$$\lambda = \varepsilon\lambda_1 + O(\varepsilon^2)$$

and note that then

$$L_0\varphi_0 + \varepsilon(L_0\varphi_1 + L_1\varphi_0) = \varepsilon\lambda_1\varphi_0 + O(\varepsilon^2)$$
$$(\|\varphi_0\|^2 - 1) + 2\varepsilon\langle\varphi_0, \varphi_1\rangle = O(\varepsilon^2)$$
$$\langle \mathbb{1}, \varphi_0\rangle + \varepsilon\langle \mathbb{1}, \varphi_1\rangle = O(\varepsilon^2).$$

Equating $O(1)$ terms,

$$L_0\varphi_0 = 0, \quad \|\varphi_0\|^2 = 1, \quad \langle \mathbb{1}, \varphi_0\rangle = 0.$$

Thus $\varphi_0 = (a\mathbb{1}^\top, b\mathbb{1}^\top)$ by Theorem 1.11 and

$$a^2N^+ + b^2N^- = 1$$
$$aN^+ + bN^- = 0.$$

Equating $O(\varepsilon)$ terms,

$$L_0\varphi_1 = \lambda_1\varphi_0 - L_1\varphi_0$$
$$\langle\varphi_0, \varphi_1\rangle = \langle \mathbb{1}, \varphi_1\rangle = 0.$$

For a solution φ_1 of the above equation to exist, note that, since $L_0\mathbb{1} = L_0\varphi_0 = 0$, invoking Fredholm alternative we require

$$\langle \mathbb{1}, \lambda_1\varphi_0 - L_1\varphi_0\rangle = 0 \tag{1.1}$$
$$\langle\varphi_0, \lambda_1\varphi_0 - L_1\varphi_0\rangle = 0 \tag{1.2}$$

and such a solution will be made unique by imposing $\langle\varphi_0, \varphi_1\rangle = \langle \mathbb{1}, \varphi_1\rangle = 0$. (1.1) is automatic since $\langle \mathbb{1}, \varphi_0\rangle = 0$ and $L_1\mathbb{1} = 0$. (1.2) is satisfied if $\lambda_1 = \langle\varphi_0, L_1\varphi_0\rangle > 0$. □

The justification for the existence and form of the power series expansions adopted in the preceding proof follows from application of the implicit function theorem to the nonlinear equations defining the eigenvalue problem.

1.4 What Ideas Did We Need?

(i) Metric spaces (Lecture 4).
(ii) Eigenvalues, eigenvectors (Lecture 13).
(iii) Orthogonality (Lecture 11).
(iv) Implicit Function Theorem (Lecture 19).

Exercises

1.1 Show that any Hermitian, positive definite matrix $A \in \mathbb{C}^{n\times n}$ can be written in the form

$$A = \begin{pmatrix} a_{11} & z^* \\ z & B \end{pmatrix} = \begin{pmatrix} c & 0 \\ \frac{1}{c}z & I \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & B - \frac{1}{a_{11}} zz^* \end{pmatrix} \begin{pmatrix} c & \frac{1}{c}z^* \\ 0 & I \end{pmatrix}$$

for some $c > 0$.

1.2 Given a Hermitian matrix $A \in \mathbb{C}^{n\times n}$ and $E \subseteq \mathbb{C}^n$ with $\dim E = k \leq n$, define the *partial trace* of A restricted to E by

$$\operatorname{tr}(A|_E) := \sum_{i=1}^{k} \langle v_i, Av_i \rangle,$$

where $\{v_1, \ldots, v_k\}$ is any orthonormal basis of E. Show that $\operatorname{tr}(A|_E)$ does not depend on this choice of basis, that it is a linear function of A, and that we have the following formulae for the eigenvalues $\lambda_1 \geq \cdots \geq \lambda_n$ of A:

$$\lambda_1 + \cdots + \lambda_k = \sup_{\substack{E\subseteq\mathbb{C}^n \\ \dim E = k}} \operatorname{tr}(A|_E) \tag{1.3}$$

and

$$\lambda_{n-k+1} + \cdots + \lambda_n = \inf_{\substack{E\subseteq\mathbb{C}^n \\ \dim E = k}} \operatorname{tr}(A|_E).$$

1.3 Using Ex. 1.2, prove the *Ky Fan inequality*: for Hermitian matrices $A, B \in \mathbb{C}^{n\times n}$, and any $1 \leq k \leq n$,

$$\sum_{i=1}^{k} \lambda_i(A + B) \leq \sum_{i=1}^{k} \lambda_i(A) + \sum_{i=1}^{k} \lambda_i(B).$$

1.4 Let $\alpha_1, \ldots, \alpha_n \in \mathbb{R}$. Define the matrix $V_n \in \mathbb{R}^{n\times n}$ by

$$V_n = \begin{pmatrix} 1 & \alpha_1 & \alpha_1^2 & \ldots & \alpha_1^{n-1} \\ 1 & \alpha_2 & \alpha_2^2 & \ldots & \alpha_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \alpha_n & \alpha_n^2 & \ldots & \alpha_n^{n-1} \end{pmatrix}.$$

(a) Show that

$$\det(V_n) = \prod_{1\leq i<j\leq n} (\alpha_j - \alpha_i).$$

(b) Let $f : \mathbb{R} \to \mathbb{R}$. Suppose that we wish to find a real polynomial P of degree at most $n - 1$,

$$P(x) = \sum_{k=0}^{n-1} u_k x^k,$$

such that $P(\alpha_j) = f(\alpha_j)$ for each $j = 1, \ldots, n$. How could such a polynomial be found, and when is this polynomial unique?

1.5 Let $G = (Z, E)$ be an undirected graph with $|Z| = N$ nodes (vertices) and let $W \in \mathbb{R}^{N \times N}$ be the corresponding (possibly weighted) adjacency matrix. Recall the definition of the unnormalized graph Laplacian

$$L = D - W$$

and the (symmetric) normalized graph Laplacian

$$L' = D^{-\frac{1}{2}} L D^{-\frac{1}{2}} = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$$

where $D \in \mathbb{R}^{N \times N}$ is the (diagonal) degree matrix. Show that L' is symmetric, positive semi-definite and has a zero eigenvalue with eigenvector $D^{\frac{1}{2}}\mathbb{1}$. **Hint**: recall properties of L.

1.6 In this problem, we focus on the unnormalized graph Laplacian L and generalize results from Lecture 1 (Theorems 1.6 and 1.11).

Given two vertices $x_i, x_j \in Z$, we will say that there exists a path between x_i and x_j if there is a set of vertices $x_{m_1}, \ldots, x_{m_p} \in Z$ such that the weights

$$W_{im_1}, W_{m_1 m_2}, W_{m_2 m_3}, \ldots, W_{m_{p-1} m_p}, W_{m_p j}$$

are all strictly positive. We will say that $A \subseteq Z$ is a connected component of G if both $x_i, x_j \in A$ implies that there exists a path between x_i and x_j, and if only one of $x_i, x_j \in A$ implies that there does not exist a path between x_i and x_j. We will say that G has k connected components if there is a disjoint collection of connected components $A_1, \ldots, A_k \subseteq G$ whose union is G.

(i) Consider Figure 1.6, representing a graph with 5 vertices. How many connected components does the graph have? What are they?

(ii) Let $L_0 \in \mathbb{R}^{(n_1+n_2)\times(n_1+n_2)}$ have block diagonal form

$$L_0 = \begin{pmatrix} L^+ & 0 \\ 0 & L^- \end{pmatrix}$$

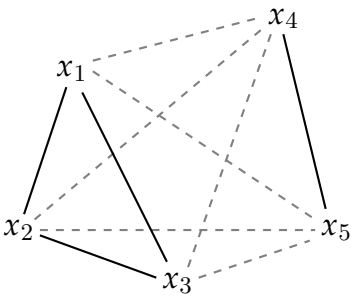


Figure 1.1 An example of a graph with weighted edges. A dashed line between x_i and x_j corresponds to $W_{ij} = 0$; a solid line corresponds to $W_{ij} > 0$.

for $L^+ \in \mathbb{R}^{n_1 \times n_1}$ and $L^- \in \mathbb{R}^{n_2 \times n_2}$. Assume that L^+ has eigenvalues $\lambda_1^1, \ldots \lambda_{n_1}^1$ with corresponding eigenvectors $v_1^1, \ldots, v_{n_1}^1 \in \mathbb{R}^{n_1}$, and L^- has eigenvalues $\lambda_1^2, \ldots \lambda_{n_2}^2$ with corresponding eigenvectors $v_1^2, \ldots, v_{n_2}^2 \in \mathbb{R}^{n_2}$. What are the eigenvalues and eigenvectors of L_0? How does this generalize to a larger number of blocks?

(iii) Show that the 0 eigenvalue of L has geometric multiplicity 1 if and only if the graph G has one connected component.

(iv) Show that the 0 eigenvalue of L has geometric multiplicity k if and only if the graph G has k connected components $\{A_m\}_{m=1}^k$. Show that in this case the eigenspace of the 0 eigenvalue is spanned by the indicator vectors $\mathbb{1}_{A_1}, \ldots, \mathbb{1}_{A_k}$.
Hint: without loss of generality you can assume that the data points are ordered according to their connected component.

This exercise tells us that when the similarity graph G has k connected components, these components can be perfectly identified given knowledge of the first k eigenvectors of the graph Laplacian L. With a good choice of weights W_{ij}, these connected components in G should correspond to the clusters in $\mathbb{R}^d$. Choice of such weights is not trivial however, as will be seen below. Nonetheless, even with imperfect weights this provides useful heuristics for determining the clusters.

1.7 *For this problem, use the Voting Records dataset from the Spectral Clustering supplement. You can use either* `voting_records.mat` *for MATLAB code or* `voting_records.zip` *with* `.csv` *files for any other language; they contain the same dataset.*

The dataset consists of the voting records of the $N = 435$ members of the U.S. House of Representatives in 1984 during the 98th U.S. Congress, 2nd session. Each record is on $d = 16$ bills with $+1$ corresponding to for, -1 to against, and 0 to abstain. Each Democrat is labeled $+1$ and each Republican -1.

Let $\mathcal{S} = \{x^{(j)}\}_{j=1}^{N}$ be a labeled dataset with corresponding labels $\{y^{(j)}\}_{j=1}^{N}$ where $x^{(j)} \in \mathbb{R}^d$ and $y^{(j)} \in \{\pm 1\}$ for $j = 1, 2, \ldots, N$. We will build a graph G using this dataset in the following way. Associate the index of each point in $\mathcal{S}$ to a vertex of the graph, namely $Z = \{1, 2, \ldots, N\}$. Construct the edge set E (or equivalently the adjacency matrix W) via a weight function $\eta\colon \mathbb{R}^+ \to \mathbb{R}^+$ (recall definition) by setting

$$W_{jk} = \eta(\|x^{(j)} - x^{(k)}\|_2) \quad \forall (j,k) \in E.$$

Use the following choices:

(1) **(ϵ-neighborhood graph)** Given some $\epsilon > 0$ define

$$\eta(r) = \begin{cases} 1, & r < \epsilon \\ 0, & \text{otherwise} \end{cases}.$$

Note that the resulting graph is unweighted and, for sufficiently small choices of ϵ, the graph is not fully-connected (why?).

(2) **(RBF kernel)** Given some $\sigma > 0$ define

$$\eta(r) = \exp\left(-\frac{r^2}{2\sigma^2}\right).$$

Note that the resulting graph is always fully-connected (why?). In machine learning, this choice of η is called the radial basis function (RBF) kernel and is frequently used in support vector machines (SVM). Other names include Gaussian kernel or squared-exponential kernel.

(3) **(Laplace kernel)** Given some $\sigma > 0$ define

$$\eta(r) = \exp\left(-\frac{r}{\sigma}\right).$$

Note again that the resulting graph is always fully-connected (why?). The name "Laplace" comes from the fact that this η resembles the (unnormalized) probability density function of a Laplace distribution. (It is NOT related to the graph Laplacian).

(4) **(Student's choice)** Construct your own weighting function η that depends on some parameter e.g. ϵ, σ. Show that your η satisfies the definition of a weight function. Is the resulting graph fully-connected? Is it weighted?

For each of the four choices above:

(a) Compute the graph Laplacians L, L' (do not print them for grading).

(b) Plot (and submit for grading) the spectra of both L, L' for different orders of the parameter in ascending order (for example, $O\left(10^s\right), s \in \{-3, \ldots, 1\}$, but don't feel constrained to these).

(c) Describe what effects you notice when changing the parameters, and what happens to the spectra of each Laplacian. Compare all four.

Fix parameter σ and describe how the spectrum changes for kernels 2 and 3 (and 4, if it is of the same structure). Again, for each of the four choices, study the first few eigenvectors (corresponding to a non-zero eigenvalue) by plotting them. Note that, as provided, the dataset is ordered, so the first $1, \ldots, K$ points belong to the first class, while the last $K+1, \ldots, N$ points belong to the second class; this makes it easy to visualize whether an eigenvector separates the data well (why?).

Come up with a definition of classification accuracy for the case of two clusters; by "definition of classification accuracy", we mean a function mapping the eigenvectors of the graph Laplacian to a real number. Note that we are doing unsupervised learning hence the ground truth labels $y^{(j)}$ are only here to see how well our algorithm does (they would not be available in practice). (Hint: the graph Laplacian has no knowledge of the ground truth i.e. it doesn't know which cluster is given $+1$ and which -1). Using your definition, decide which kernel works best and whether the normalized or unnormalized graph Laplacian should be preferred. Can you tune the kernel parameters such that each graph Laplacian separates the data?

1.8 *For this problem, use the Synthetic dataset from the Spectral Clustering supplement. You can use either* `synthetic.mat` *for MATLAB code or* `synthetic.zip` *with* `.csv` *files for any other language; they contain the same dataset.*

The dataset consists of three sets: `X1`, `X2`, `X3`; they arise from a

number of points distributed into a number of clusters in the plane, embedded in $\mathbb{R}^{100}$ and then perturbed by noise in all directions. Each set has the form of a matrix $X \in \mathbb{R}^{N\times d}$, with each row of X representing a data point $x_i \in \mathbb{R}^d$.

In this problem, we will only consider the following weighted adjacency matrix:

$$W_{ij} = \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|_2^2}{\ell^2}\right),$$

i.e. the RBF kernel with $\sigma = \ell/\sqrt{2}$.

(i) Compute the eigenvectors of the unnormalized Laplacians for a variety of choices of $\ell > 0$, using the data `X1`. From Lecture 1 you know that information about clusters should lie in the first few eigenvectors — by looking at these, attempt to infer the number of clusters and their elements. How sensitive is the clustering to the choice of ℓ?

(ii) Instead of a fixed length-scale parameter $\ell > 0$, the length scale can be inferred from the data. Fix $K \in \mathbb{N}$. The self-tuning weights of Zelnik-Manor and Perona (2004) are defined by

$$W_{ij} = \exp\left(-\frac{\|x^{(i)} - x^{(j)}\|_2^2}{\ell_i \ell_j}\right) \tag{1.4}$$

where $\ell_j = \|x^{(j)} - x_K\|_2$, and x_K is the K-th nearest neighbor of x_j. An example MATLAB implementation of these weights is provided in `weights_st.m`. Using this, or your own implementation if you wish, repeat the clustering from part (i) using the self-tuning weights with $K = 10$.

(iii) The three sets of data `X1`, `X2`, `X3` arise from the same clusters, perturbed by increasing levels of noise. How does the noise level affect the ability to determine the clusters?

(iv) Illustrate the output of some of the above clustering via a projection of the clustered data onto its first two principle components.

2
Control Motivation

This lecture demonstrates the role of linear analysis in the formulation of problems arising in control theory. The linear operator mapping initial condition and controls to the solution trajectory in state space is studied, in the context of both discrete and continuous time linear control problems.

2.1 Discrete Time Linear Control

2.1.1 Background Material

We use the notation $\mathbb{Z}^+ = \{0, 1, 2, \dots\}$ and $\mathbb{N} = \{1, 2, 3, \dots\}$. Let $x_k \in \mathbb{R}^n$, $u_k \in \mathbb{R}^m$ for $k \in \mathbb{Z}^+$, $a \in \mathbb{R}^n$ and fix $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. Consider the iteration

$$x_{k+1} = Ax_k + Bu_k, \quad x_0 = a.$$

Definition 2.1. *The controllable set C is the set of $a \in \mathbb{R}^n$ for which $\exists l \in \mathbb{Z}^+$ and $\{u_j\}_{j=0}^{l-1}$ such that $x_l = 0$. If $C = \mathbb{R}^n$, then the system is controllable.*

Definition 2.2. *The controllability matrix is*

$$G = (B, AB, A^2B, \dots, A^{n-1}B) \in \mathbb{R}^{n\times nm}.$$

Theorem 2.3. *Assume A is invertible. Then the system is controllable if and only if* $\operatorname{rank} G = n$.

In this lecture our focus is on studying the map from initial condition a and control $\{u_k\}$ to the solution $\{x_k\}$. To this end the following lemma is useful:

Lemma 2.4. $x_k = A^k a + \sum_{j=0}^{k-1} A^{k-1-j} B u_j$.

Proof We proceed by induction. The statement is clearly true for $k = 0$. Assume it is true for k. Then

$$\begin{aligned} x_{k+1} &= A\left(A^k a + \sum_{j=0}^{k-1} A^{k-1-j} B u_j\right) + B u_k \\ &= A^{k+1} a + \sum_{j=0}^{k-1} A^{k-j} B u_j + B u_k \\ &= A^{k+1} a + \sum_{j=0}^{k} A^{k-j} B u_j \end{aligned}$$

so it is true for $k + 1$. □

In what follows we will make the following assumptions:

Assumption 2.5. *We assume that there are norms $\|\cdot\|$ on matrices and vectors such that, for all $v \in \mathbb{R}^q$, $M \in \mathbb{R}^{p\times q}$,*

$$\|Mv\| \le \|M\|\|v\|.$$

Then assume that in the case $p = q = n$ there exists $\alpha \in (0, 1)$ such that

$$\|A^k\| \le \alpha^k.$$

Note that the first assumption is an assumption about a *general* property of norms on matrices and vectors; the second is a property of the *specific* matrix A in such a norm.

Theorem 2.6. *Let Assumption 2.5 hold. Then, for all $k \in \mathbb{Z}^+$,*

$$\|x_k\| \le \|a\| + \|B\|\left(\sum_{j=0}^{k-1} \alpha^j\right) \max_{0\le j\le k-1} \|u_j\|.$$

Proof We have, using Lemma 2.4 and Assumption 2.5,

$$\begin{aligned} \|x_k\| &\le \|A^k\|\|a\| + \sum_{j=0}^{k-1} \|A^{k-1-j}\|\|Bu_j\| \\ &\le \alpha^k \|a\| + \sum_{j=0}^{k-1} \alpha^{k-1-j} \|B\|\|u_j\| \\ &\le \|a\| + \|B\|\left(\sum_{j=0}^{k-1} \alpha^j\right) \max_{0\le j\le k-1} \|u_j\|. \end{aligned}$$

□

2.1.2 Finite Dimensional Setting

Let $x = \{x_k\}_{k=0}^{K}$, $w = \{a, \{u_k\}_{k=0}^{K-1}\}$, and

$$\mathsf{A} = \begin{bmatrix} I & & & & \\ -A & I & & & \\ & -A & I & & \\ & & \ddots & & \\ & & & -A & I \end{bmatrix} \qquad \mathsf{B} = \begin{bmatrix} I & & & & \\ & B & & & \\ & & B & & \\ & & & \ddots & \\ & & & & B \end{bmatrix}$$

where $I \in \mathbb{R}^{n\times n}$ and, recall, $A \in \mathbb{R}^{n\times n}$, and $B \in \mathbb{R}^{n\times m}$. Then

$$\mathsf{A}x = \mathsf{B}w.$$

Note that $\mathsf{A} \in \mathbb{R}^{(K+1)n\times(K+1)n}$ and $\mathsf{B} \in \mathbb{R}^{(K+1)n\times(n+Km)}$. This is thus a finite system of linear equations to determine x from w. Lemma 2.4 shows that the linear system has a solution, and it is left as an exercise to show that the solution is unique. Thus we have shown that A is invertible and we may define $\mathsf{L} = \mathsf{A}^{-1}\mathsf{B}$. It follows that $x = \mathsf{L}w$.

Theorem 2.6 shows that, for all $k \in \{0, \cdots, K\}$,

$$\|x_k\| \le \|a\| + \|B\| \left(\sum_{j=0}^{\infty} \alpha^j \right) \max_{0\le j\le K-1} \|u_j\|.$$

Thus

$$\max_{0\le j\le K} \|x_j\| \le \|a\| + \frac{1}{1-\alpha}\|B\| \max_{0\le j\le K-1} \|u_j\|.$$

If we define

$$\|x\| := \max_{0\le j\le K} \|x_j\| \quad \& \quad \|w\| := \max\Big\{\|a\|, \max_{0\le j\le K-1} \|u_j\|\Big\}$$

(these are valid norms), then we have shown that

$$\|x\| \le C\|w\|$$

where

$$C = 1 + \frac{1}{1-\alpha}\|B\|.$$

Since $x = \mathsf{L}w$ and since w is an arbitrary vector (of appropriate dimensions) we have shown that

$$\|\mathsf{L}\| = \sup_{w \neq 0} \frac{\|\mathsf{L}w\|}{\|w\|} \leq C.$$

This interesting property measures the size of the linear map L from the data w to the solution x. We say that L is a *bounded linear operator.* In fact L is a (finite) matrix and all (finite) matrices are bounded linear operators.

Furthermore, although we used very specific norms on x and w, and hence on L, in this finite dimensional setting different norms would only change the value of the constant C; it will always be finite, no matter what norm is chosen. This is because of *norm equivalence* in finite dimensions. The situation is very different in infinite dimensions.

2.1.3 Infinite Dimensional Setting

We now proceed to solve for the entire sequence $\{x_k\}_{k\in\mathbb{Z}^+}$. To this end let $x = \{x_k\}_{k\in\mathbb{Z}^+}$, $w = \{a, \{u_k\}_{k\in\mathbb{Z}^+}\}$, and

$$\mathsf{A} = \begin{bmatrix} I & & & \\ -A & I & & \\ & -A & I & \\ & & & \ddots \end{bmatrix} \quad \mathsf{B} = \begin{bmatrix} I & & & \\ & B & & \\ & & B & \\ & & & \ddots \end{bmatrix}$$

where $I \in \mathbb{R}^{n\times n}$, $A \in \mathbb{R}^{n\times n}$, and $B \in \mathbb{R}^{n\times m}$. Then the original iteration may be formulated as

$$\mathsf{A}x = \mathsf{B}w.$$

This is a countably infinite system of linear equations and the matrices A, B are now infinite matrices. A number of natural questions arise. Does the equation have a solution? If so, in what sense? Can we find L such that $x = \mathsf{L}w$? To answer this question we need to define spaces $\mathcal{U}$ and $\mathcal{X}$ and define $\mathsf{L}\colon \mathcal{U} \to \mathcal{X}$.

Definition 2.7. *Let $v = \{v_k\}_{k\in\mathbb{Z}^+}$, $v_k \in \mathbb{R}^l$. Given $w = \{w_k\}_{k\in\mathbb{Z}^+}$, $w_k \in (0,\infty)$ and $p \in [1,\infty)$, we define*

$$\ell^p_w(\mathbb{Z}^+;\mathbb{R}^l) = \{v : \|v\|_{\ell^p_w} < \infty\}$$

where

$$\|v\|_{\ell^p_w} = \Big(\sum_{k\in\mathbb{Z}^+} w_k \|v_k\|^p\Big)^{1/p}$$

for $\|\cdot\|$ any norm on $\mathbb{R}^l$. If $w_k \equiv 1$, we write $\ell^p(\mathbb{Z}^+;\mathbb{R})$, $\|\cdot\|_{\ell^p}$. If $p=\infty$, we define

$$\ell^\infty(\mathbb{Z}^+;\mathbb{R}^l) = \{v : \|v\|_{\ell^\infty} < \infty\}$$

where

$$\|v\|_{\ell^\infty} = \sup_{k\in\mathbb{Z}^+} \|v_k\|$$

for $\|\cdot\|$ any norm on $\mathbb{R}^l$.

Theorem 2.8. *Under Assumption 2.5,* A *is invertible and there is a bounded linear operator* L *mapping $w \in \mathcal{U} := \mathbb{R}^n \times \ell^\infty(\mathbb{Z}^+;\mathbb{R}^m)$ into $x \in \mathcal{X} := \ell^\infty(\mathbb{Z}^+;\mathbb{R}^n)$.*

Proof To establish the boundedness of the mapping from w to x in the specified spaces, note that we have

$$\begin{aligned}
\|x_k\| &\le \|A^k\|\|a\| + \sum_{j=0}^{k-1} \|A^{k-1-j}\|\|Bu_j\| \\
&\le \alpha^k\|a\| + \sum_{j=0}^{k-1} \alpha^{k-1-j}\|B\|\|u_j\| \\
&\le \|a\| + \|B\|\left(\sum_{j=0}^{k-1}\alpha^j\right)\sup_j\|u_j\| \\
&\le \|a\| + \frac{1}{1-\alpha}\|B\|\|u\|_{\ell^\infty}.
\end{aligned}$$

Hence

$$\|x\|_{\ell^\infty} = \sup_k \|x_k\| \le \|a\| + \frac{1}{1-\alpha}\|B\|\|u\|_{\ell^\infty}$$

If we define

$$\|x\|_{\mathcal{X}} := \|x\|_{\ell^\infty} \quad \& \quad \|w\|_{\mathcal{U}} = \max\{\|a\|, \|u\|_{\ell^\infty}\}$$

(these are valid norms) then we have shown that, for the same constant C from the finite dimensional setting,

$$\|x\|_{\mathcal{X}} \le C\|w\|_{\mathcal{U}}. \tag{2.1}$$

The existence of a map from w to x is proven by the constructive formula of Lemma 2.4, which holds for all $k \in \mathbb{Z}^+$. The linearity of the

mapping may be established as follows. Within the aforementioned constructive formula, replace $(a, u) \mapsto (\theta a^{(1)} + \phi a^{(2)}, \theta u^{(1)} + \phi u^{(2)})$. Then

$$x_k = A^k(\theta a^{(1)} + \phi a^{(2)}) + \sum_{j=0}^{k-1} A^{k-1-j} B(\theta u_j^{(1)} + \phi u_j^{(2)})$$
$$= \theta \left(A^k a^{(1)} + \sum_{j=0}^{k-1} A^{k-1-j} B u_j^{(1)} \right) + \phi \left(A^k a^{(2)} \sum_{j=0}^{k-1} A^{k-1-j} B u_j^{(2)} \right).$$

Both terms in parentheses are finite if $u^{(1)}, u^{(2)} \in \ell^\infty(\mathbb{Z}^+; \mathbb{R}^m)$, respectively, and $x \in \ell^\infty(\mathbb{Z}^+; \mathbb{R}^n)$ is defined as

$$x = \theta \mathsf{L} w^{(1)} + \phi \mathsf{L} w^{(2)}, \quad w^{(i)} = \begin{bmatrix} a^{(i)} \\ u^{(i)} \end{bmatrix}.$$

Hence, we have established linearity. Uniqueness of x, given w, is established by assuming the existence of two solutions x and y, for the same w, and noting that

$$x_{k+1} = Ax_k + Bu_k, \quad x_0 = a,$$
$$y_{k+1} = Ay_k + Bu_k, \quad y_0 = a$$

and subtracting shows that $e_k = x_k - y_k$ satisfies

$$e_{k+1} = Ae_k, \quad e_0 = 0.$$

It follows that $e_k = 0$ for all $k \in \mathbb{Z}^+$ and hence that $x_k = y_k$ for all $k \in \mathbb{Z}^+$ as required for uniqueness. Thus we may write $x = \mathsf{L}w$ for $w = (a, u)$.

From (2.1) we have, since $w \in \mathcal{U}$ is arbitrary,

$$\|\mathsf{L}\| = \sup_{w \neq 0} \frac{\|\mathsf{L}w\|_{\mathcal{X}}}{\|w\|_{\mathcal{U}}} \leq C.$$

Thus L is a bounded linear operator, as required.

□

2.2 Continuous Time Linear Control

2.2.1 Background Material

We consider the differential equation

$$\frac{dx}{dt}(t) = Ax(t) + Bu(t), \quad x(0) = a$$

for $t \in (0, \infty)$ with $u \in C([0, \infty); \mathbb{R}^m)$ and $x \in C([0, \infty); \mathbb{R}^n)$.

Definition 2.9. *The controllable set C is the set of $a \in \mathbb{R}^n$ for which there is a $T > 0$ and $u \in C([0,T];\mathbb{R}^m)$ such that $x(T) = 0$. If $C = \mathbb{R}^n$, then the system is said to be controllable.*

Theorem 2.10. *The system is controllable if and only if* $\operatorname{rank} G = n$.

2.2.2 Infinite Dimensional Setting

Lemma 2.11. $x(t) = e^{At}a + \int_0^t e^{A(t-s)}Bu(s)ds$.

Proof We have

$$e^{-At}\frac{dx}{dt}(t) - Ae^{-At}x(t) = e^{-At}Bu(t),$$

hence

$$\frac{d}{dt}\left(e^{-At}x(t)\right) = e^{-At}Bu(t).$$

Integrating gives

$$e^{-At}x(t) - a = \int_0^t e^{-As}Bu(s)ds$$

and thus

$$x(t) = e^{At}a + \int_0^t e^{A(t-s)}Bu(s)ds.$$

□

Assumption 2.12. *We assume that there are norms $\|\cdot\|$ on matrices and vectors such that, for all $v \in \mathbb{R}^n$, $M \in \mathbb{R}^{n\times n}$,*

$$\|Mv\| \le \|M\|\|v\|$$

and there is a $c \in \mathbb{R}^+$, $\lambda > 0$ for which, in this matrix norm,

$$\|e^{At}\| \le ce^{-\lambda t}.$$

Let $x = \{x(t)\}_{t\ge 0}$ and $w = \{a, \{u(t)\}_{t\ge 0}\}$.

Definition 2.13. *For $v\colon \mathbb{R}^+ \to \mathbb{R}^l$ and $p \in [1,\infty)$, define*

$$L^p(\mathbb{R}^+;\mathbb{R}^l) = \{v : \|v\|_{L^p} < \infty\}$$

where

$$\|v\|_{L^p}^p = \int_0^\infty \|v(s)\|^p ds$$

for $\|\cdot\|$ *any norm on* $\mathbb{R}^l$. *For* $p = \infty$, *define*

$$L^\infty(\mathbb{R}^+;\mathbb{R}^l) = \{v : \|v\|_{L^\infty} < \infty\}$$

where

$$\|v\|_{L^\infty} = \sup_{t\geq 0} \|v(t)\|$$

for $\|\cdot\|$ *any norm on* $\mathbb{R}^l$.

Theorem 2.14. *Under Assumptions 2.12,* $x = \mathsf{L}w$, *where* L *is a linear operator from* $\mathcal{U} := \mathbb{R}^n \times L^\infty(\mathbb{R}^+;\mathbb{R}^m)$ *into* $\mathcal{X} := L^\infty(\mathbb{R}^+;\mathbb{R}^n)$.

Proof We have

$$\begin{aligned}\|x(t)\| &\leq ce^{-\lambda t}\|a\| + c\int_0^t e^{-\lambda(t-s)}ds\|B\|\|u\|_{L^\infty} \\ &\leq c\|a\| + \frac{c}{\lambda}\|B\|\|u\|_{L^\infty}.\end{aligned}$$

Hence

$$\|x\|_{L^\infty} = \sup_{t\geq 0}\|x(t)\| \leq c\|a\| + \frac{c}{\lambda}\|B\|\|u\|_{L^\infty}.$$

This establishes boundedness; linearity may be proved similarly to the discrete time case. □

2.3 Further Discussion of Continuous Time Control Theory

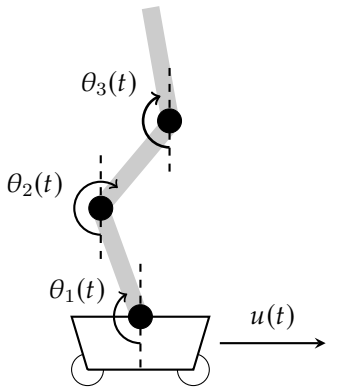
Systems of ordinary differential equations (ODEs) can be used to describe many physical processes, from the dynamics of populations to the trajectories of missiles. The general form for such systems is as follows:

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = F(x(t),t), \quad x(0) = x_0,$$

where $x(t) \in \mathbb{R}^n$ for each $t \geq 0$. Note that a higher order system can always be written as a first order system in a higher dimensional space. The behavior of the solution $x(t)$ may be undesirable, and so it will often be of interest to control the system by, for example, forcing it. The above system is therefore generalized to

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = F(x(t),t,u(t)), \quad x(0) = x_0,$$

where $u : [0, \infty) \to \mathbb{R}$ is a control function. The function u may also depend on the state $x(t)$ at time t, in addition to t, to allow for feedback into the system. The question arises of how the control u can be chosen to enforce certain behavior of the solution x.



The triple pendulum on a cart.

As an example, consider a triple pendulum on a cart as illustrated in Figure 2.3. The state of the pendulum can be characterized by the pivot angles $(\theta_1, \theta_2, \theta_3)$. Under the laws of classical mechanics, the evolution of the angles $(\theta_1, \theta_2, \theta_3)$ in time is deterministic, though their behavior is chaotic and the system of differential equations that describes their evolution is nonlinear. It may be of interest to ask how the cart should be moved, according to the control $u(t)$, to ensure that each $\theta_i(t) \to \pi$ in a finite time so that the pendulum stands up straight. The following video illustrates the calculation and implementation of such a control:

https://www.youtube.com/watch?v=cyN-CRNrb3E

We define the space of *unrestricted controls* $\mathcal{U}$ as the set of functions $u : [0, \infty) \to \mathbb{R}^m$:

$$\mathcal{U} = \{u : [0, \infty) \to \mathbb{R}^m\}.$$

It is often of interest in control theory to work with a subset $\mathcal{U}_{ad} \subset \mathcal{U}$, referred to as the set of *admissible controls*, however we consider only unrestricted controls.

Let $A \in \mathbb{R}^{n \times n}$ and $B \in \mathbb{R}^{n \times m}$. Given a control $u \in \mathcal{U}$, consider the *response* $x : [0, \infty) \to \mathbb{R}^n$ defined via the autonomous system

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = Ax(t) + Bu(t), \quad x(0) = x_0 \tag{2.2}$$

for some $x_0 \in \mathbb{R}^n$. If $u(\cdot)$ depends only upon the initial condition x_0, it is called an *open-loop control*. If it depends on the path $x(\cdot)$, it it called a *closed-loop control*. A closed-loop control allows feedback into the system, whereas an open-loop control does not.

We introduce the notion of *controllability* for this system:

Definition 2.15 (Controllability)**.** *For each time $t > 0$, the* fixed time t controllable set *is defined by*

$$C(t) = \{x_0 \in \mathbb{R}^n : \text{there exists } u \in \mathcal{U} \text{ with } x(t) = 0\}.$$

The controllable set *is defined by*

$$C = \bigcup_{t>0} C(t).$$

If $C = \mathbb{R}^n$, we will say that the system is controllable*. We will alternatively say that (A, B) is controllable in this case.*

The system is hence controllable if it can be controlled to hit zero in a finite time, from any starting point. A useful characterization of controllability is given in terms of the *controllability matrix* $G(A, B) \in \mathbb{R}^{n\times mn}$ of (2.2). This is defined by

$$G(A, B) = \big(B, AB, A^2B, \dots, A^{n-1}B\big).$$

Theorem 2.16. *(A, B) is controllable if and only if* $\operatorname{rank}\big(G(A, B)\big) = n$.

Proof of this theorem may be found, for example, in Evans (2005).

We now introduce *partially observed systems*. Suppose that we wish to recover a signal $x(t)$ that is known to satisfy the system

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = Ax(t), \quad x(0) = x_0,$$

but we do not know the initial condition x_0. We do not observe $x(t)$ directly, but instead have observations given by

$$y(t) = Hx(t)$$

for some $H \in \mathbb{R}^{m\times n}$. We wish to use these observations to determine $x(t)$ for sufficiently large t. To this end, we consider the system given by

$$\frac{\mathrm{d}z}{\mathrm{d}t}(t) = Az(t) + B\big(y(t) - Hz(t)\big), \quad z(0) = z_0 \tag{2.3}$$

for some $B \in \mathbb{R}^{n\times m}$. This is of the form (2.2) with the closed loop control $u = y - Hz$. The reasoning for this is that when z is close to x, so that y is close to Hz, the dynamics of z should be close to those of x. In particular, if $z(t_*) = x(t_*)$ for some $t_* \geq 0$, then $z(t) = x(t)$ for all $t \geq t^*$. In general, we cannot guarantee that the positions of z and x will ever coincide, but we can instead aim for $z(t)$ and $x(t)$ to become arbitrarily close for large times t.

In Exercise 2.4, we show how solutions to general linear systems of ODEs can be found via use of the matrix exponential, the concept you will learn about in one of the last lectures of the course. In Exercise 2.5, we explore the notion of controllability by looking at specific examples and conditions for when it holds for linear systems. In Exercise 2.6, we consider partially observed systems and look at when we can infer system dynamics from these observations using a consequence of controllability. In Exercise 2.7, we study the linear systems from Exercise 2.6 numerically, along with the Lorenz '63 model, which is a nonlinear and chaotic system.

2.4 What Ideas Did We Need?

(i) Norm equivalence (Lecture 4).
(ii) Matrix norms (Lecture 7).
(iii) Linear operators on Banach spaces (Lecture 7).
(iv) Rank of a linear operator (Lecture 7).
(v) ℓ^p spaces (Lecture 5).
(vi) L^p spaces (Lecture 5).
(vii) e^{At} and A^k (Lecture 16).

Exercises

2.1 The Fibonacci numbers $\{F_n\}_{n\in\mathbb{N}}$ are defined by the relation

$$F_n = F_{n-1} + F_{n-2}, \quad F_0 = 0, \quad F_1 = 1.$$

(a) Write this equation in the form

$$u_n = Au_{n-1}, \quad u_0 = u$$

for $u_n, u \in \mathbb{R}^2$, $A \in \mathbb{R}^{2\times 2}$. Hence find an explicit expression for $F_n = \mathrm{Fib}(n)$, $n \in \mathbb{N}$, where $\mathrm{Fib}\colon \mathbb{N} \to \mathbb{R}$ does not depend on any of the F_k's, $k \in \mathbb{N}$.

(b) What choice of initial conditions on F_0, F_1 would give rise to an ℓ^∞ sequence F_n?

2.2 Consider an operator A on infinite real-valued sequences $x =$

$\{x_n\}_{n\in\mathbb{Z}^+}$ defined by

$$\mathsf{A}x = \begin{bmatrix} x_0 \\ x_1 - \alpha x_0 \\ x_2 - \alpha x_1 \\ \vdots \end{bmatrix}$$

for some $|\alpha| < 1$. Show that if $\mathsf{A}x = v$ and $v \in \ell^2(\mathbb{Z}^+;\mathbb{R})$, then x is uniquely defined and is an element of $\ell^\infty(\mathbb{Z}^+;\mathbb{R})$ (this shows that A is invertible when viewed as an operator from $\ell^\infty(\mathbb{Z}^+;\mathbb{R})$ to $\ell^2(\mathbb{Z}^+;\mathbb{R})$). Show explicitly that

$$\|x\|_{\ell^\infty} \leq \frac{1}{(1-\alpha^2)} + \|v\|_{\ell^2}^2.$$

2.3 Recall the discrete-time control problem

$$\mathsf{A}x = \mathsf{B}w$$

where $x = \{x_k\}_{k\in\mathbb{Z}^+}$ with $x_k \in \mathbb{R}^n$ and $w = \{a, \{u_k\}_{k\in\mathbb{Z}^+}\}$ with $a \in \mathbb{R}^n$ and $u_k \in \mathbb{R}^m$. The operators A, B are defined as the infinite block matrices

$$\mathsf{A} = \begin{bmatrix} I & & & \\ -A & I & & \\ & -A & I & \\ & & \ddots & \end{bmatrix} \quad \mathsf{B} = \begin{bmatrix} I & & & \\ & B & & \\ & & B & \\ & & & \ddots \end{bmatrix}$$

for some $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times m}$. Further recall that we may express

$$x_k = A^k a + \sum_{j=0}^{k-1} A^{k-1-j} B u_j.$$

We make the following assumptions:

- matrix A is invertible;
- matrices A and B are such that the system is controllable;
- there is a finite-dimensional norm $\|\cdot\|$ on matrices and vectors such that for any matrix M

$$\|Mv\| \leq \|M\|\|v\|;$$

- there exists $\alpha \in (0,1)$ such that

$$\|A^k\| \leq \alpha^k.$$

Finally, we introduce a *class* of inverse operators $\mathsf{L}_{A,B} : w \mapsto x$, parametrized by $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. We will simply write $\mathsf{L} \equiv \mathsf{L}_{A,B}$. Formally, $\mathsf{L} = \mathsf{A}^{-1}\mathsf{B}$.

(a) Show that the inverse operator L, viewed as an operator between following spaces:

$$\mathsf{L} : \mathbb{R}^n \times \ell^2(\mathbb{Z}^+; \mathbb{R}^m) \to \ell^\infty(\mathbb{Z}^+; \mathbb{R}^n),$$

is bounded and linear for any A and B satisfying assumptions above.

(b) Show that the inverse operator L, when viewed as

$$\mathsf{L} : \mathbb{R}^n \times \ell^\infty(\mathbb{Z}^+; \mathbb{R}^m) \to \ell^2(\mathbb{Z}^+; \mathbb{R}^n),$$

is NOT a bounded linear operator for any A, B satisfying assumptions above.

Hint: construct a counterexample; however, note that neither finding explicit matrices A and/or B nor considering special cases ($n = 1$ or $m = 1$, or both) would constitute a proper counterexample. The operator family L parametrized by matrices A and B should be left in general form; you may only use the assumptions made about the matrices.

2.4 *This is the first of four exercises on continuous time control theory. See Section 2.3 for reference and definitions.*

(a) Let $A \in \mathbb{R}^{n\times n}$. Define the matrix e^A, which we call *exponential of* A, by the Taylor series

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k. \tag{2.4}$$

In Lectures 15 & 16, we will develop a rigorous approach to functions of matrices in general, but for now you may treat (2.4) as a formal definition.

Show that for each $t \geq 0$,

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(e^{tA}\right) = Ae^{tA} = e^{tA}A,$$

assuming that all matrix exponentials exist and that $e^{A+B} = e^A e^B$ if $AB = BA$.

(b) Let $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$.

(i) Using part (a), write down the solution to the linear system

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = Ax(t), \quad x(0) = x_0. \tag{2.5}$$

Hint: try guessing what the solution should look like, analogously to the $n = 1$ case, and then show that it solves (2.5).
If $x_0 \in \mathbb{C}^n$ is an eigenvector of A with eigenvalue $\lambda \in \mathbb{C}$, describe the behavior of $x(t)$ based on the value of λ.

(ii) Let $u : [0, \infty) \to \mathbb{R}^m$ be a control function. Consider the forced linear system given by

$$\frac{\mathrm{d}x}{\mathrm{d}t}(t) = Ax(t) + Bu(t), \quad x(0) = x_0.$$

The solution to this system is given by

$$x(t) = e^{tA}x_0 + \int_0^t e^{(t-s)A}Bu(s)\,\mathrm{d}s.$$

By use of a suitable integrating factor, derive this solution.

2.5 *This is the second of four exercises on continuous time control theory. See Section 2.3 for reference and definitions.*

(a) The Cayley–Hamilton theorem states that every square matrix satisfies its own characteristic polynomial. Using this, explain why

$$\operatorname{rank}\big(G(A,B), A^kB\big) = \operatorname{rank}\big(G(A,B)\big)$$

for any $k \geq n$.

(b) Let $n = m = 2$.

(i) Let $B = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Under what conditions on A is the associated system controllable?

(ii) Let $B = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$. Under what conditions on A is the associated system controllable?

(c) Let $n = 3$, $m = 1$, and define $A \in \mathbb{R}^{n\times n}$ by

$$A = \begin{pmatrix} 0 & 1 & 0 \\ 0 & -1 & 1 \\ 0 & 0 & -1 \end{pmatrix}.$$

Consider the three cases

$$B = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \text{ and } \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

For which of these choices of B is the associated system controllable? In the case(s) where the system is not controllable, find the controllable set C **(hard)**.
You may use `Mathematica/WolframAlpha` to compute e^{tA}.

2.6 *This is the third of four exercises on continuous time control theory. See Section 2.3 for reference and definitions.*

(a) Define the error $e(t) = z(t) - x(t)$. Show that $e(t)$ satisfies the equation

$$\frac{\mathrm{d}e}{\mathrm{d}t}(t) = (A - BH)e(t), \quad e(0) = z_0 - x_0$$

and hence give a sufficient criterion that ensures $e(t) \to 0$ as $t \to \infty$ for any choice of z_0.

(b) If (A, B) is controllable, it is known that there exists an observation matrix $H \in \mathbb{R}^{m \times n}$ such that $e(t) \to 0$.

(i) Consider the cases from Exercise 2.5(b)-(c) which were controllable. By hand, find observation matrices H such that $e(t) \to 0$.

(ii) In the cases from Exercise 2.5(b)-(c) where (A, B) were not controllable, can you find $H \in \mathbb{R}^{m \times n}$ such that $e(t) \to 0$ for any z_0? Investigate either numerically or by hand, choosing at least one example from each of 2.5(b) and (c) where controllability does not hold, and looking at the spectrum of $A - BH$ for different choices of H. You may use software of your choice to find spectra.

Remark 2.17. *Above we are given the matrices A, B, and use controllability to see that we can choose an observation matrix H that causes $e(t) \to 0$. There is also a dual notion to controllability, called* observability. *If we are instead given the matrices A, H, and the pair (A, H) is observable, then it can be shown that we can choose a matrix B such that $e(t) \to 0$.*

2.7 *This is the fourth of four exercises on continuous time control theory. See Section 2.3 for reference and definitions.*

(a) For this part, you will need to implement the systems (2.2), (2.3) for arbitrary $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$. You can use the forms of the solutions given in Exercise 2.4. Solve using one of implemented ODE solvers from a numerical library such as `ode45`, `dopri5` etc., or implement a Runge–Kutta method directly. Do not submit the code for grading, just the plots.

(i) Choose an example from Exercise 2.5(b) where you know that $e(t) \to 0$ for any z_0. Fix x_0, and verify that this is the case by plotting (on the same axes) the trajectories of $e(\cdot)$ for variety of choices of z_0 (please plot on the phase plane, not as a function of time).

(ii) Let $n = m = 3$. Define $A \in \mathbb{R}^{3\times 3}$ by

$$A = \begin{pmatrix} 0.1 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}$$

and let $B = I$. Observe that $\operatorname{rank}\big(G(A,B)\big) = 3$, and so there exists an observation matrix $H \in \mathbb{R}^{3\times 3}$ that drives the error to zero.

Fix $x_0 = (1,1,0)$ and $z_0 = (2,2,2)$. Consider the three observation matrices H given by

$$H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ and } \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}.$$

For each H, plot the error $\varepsilon(t) = \|e(t)\| = \|x(t) - z(t)\|$ for $t \in [0, 50]$. By considering the trajectories of $x(\cdot)$ and $z(\cdot)$, and the structures of A, B, H, explain the behavior of these errors.

(b) We now consider a nonlinear example. The Lorenz '63 model is defined as follows:

$$\begin{aligned} \frac{\mathrm{d}x_1}{\mathrm{d}t}(t) &= \sigma\big(x_2(t) - x_1(t)\big) \\ \frac{\mathrm{d}x_2}{\mathrm{d}t}(t) &= x_1(t)\big(\rho - x_3(t)\big) - x_2(t) \\ \frac{\mathrm{d}x_3}{\mathrm{d}t}(t) &= x_1(t)x_2(t) - \beta x_3(t) \end{aligned}$$

where σ, ρ, β are scalar parameters and $x_1(t), x_2(t), x_3(t) \in \mathbb{R}$ for each t. In what follows we will fix $\sigma = 10$, $\beta = 8/3$ and $\rho = 28$; it is known that with these choices the system is chaotic.

We can write the above system more compactly as

$$\begin{aligned} \frac{\mathrm{d}x}{\mathrm{d}t}(t) &= F(x(t)), \\ x(0) &= \big(x_1(0), x_2(0), x_3(0)\big). \end{aligned} \tag{2.6}$$

where now $F\colon \mathbb{R}^3 \to \mathbb{R}^3$ is nonlinear. Suppose that we have observations

$$y(t) = Hx(t)$$

for some $H \in \mathbb{R}^{3\times 3}$, and analogously to the linear case (2.3), consider the controlled system

$$\begin{aligned} \frac{\mathrm{d}z}{\mathrm{d}t}(t) &= F(z(t)) + B\big(y(t) - Hz(t)\big), \\ z(0) &= \big(z_1(0), z_2(0), z_3(0)\big). \end{aligned} \tag{2.7}$$

for some $B \in \mathbb{R}^{3\times 3}$.

(i) Implement system (2.6). Plot (in phase space, 3-d perspective) the solution to (2.6) for a number of close initial conditions and $t \in [0, 50]$. Observe the sensitivity of the solution to the choice of initial condition: two distinct but arbitrarily close initial conditions can lead to very different trajectories.

(ii) Implement system (2.7) for arbitrary observation and control matrices H, B. We consider the case where we only observe one component of the solution, so that $H \in \mathbb{R}^{3\times 3}$ is given by

$$H = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \begin{pmatrix} 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix} \text{ or } \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Additionally we assume $B = \alpha I$ for some $\alpha > 0$, where I is the identity matrix, and consider $t \in [0, 100]$.

Fix $x(0)$ and $z(0)$, with $\|x(0) - z(0)\| \geq 20$. Define the error $\varepsilon(t) = \|x(t) - z(t)\|$. For each choice of H above, can you find $\alpha^* > 0$ such that that $\varepsilon(t) \approx 0$ for large t and for all $\alpha > \alpha^*$? If this is the case, how small can you take α^* for this to hold? Produce a plot of the error $\varepsilon(t)$ for each choice of H, with $\varepsilon(t)$ becoming small if possible.

3
Partial Differential Equations Motivation

This lecture demonstrates the role of linear analysis in the formulation of problems arising in partial differential equations (PDEs). We start by defining scalar and vector fields, and the notions of gradient and divergence. We then show how PDEs arise in two applications: as large graph limits of the graph Laplacian and in the modeling of flow in porous media. We also exhibit the Galerkin method for approximation of the resulting PDEs.

3.1 Notation

We use D to denote a bounded open subset of $\mathbb{R}^d$ and $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$. In this section we use the following functions with domain D and various ranges (i.e., codomains):

- Scalar field $f : D \to \mathbb{R}$.
- Vector field $\varphi : D \to \mathbb{R}^d$
- Matrix field $A : D \to \mathbb{R}^{d \times d}$

We may combine scalar, vector and matrix fields in the natural ways: $A\varphi$ is a vector field, $f\varphi$ is a vector field, the composition of two matrix fields is a matrix field, the outer product of two vector fields is a matrix field, inner product of two vector fields is a scalar field, the product of two scalar fields is a scalar field, and so forth.

We may now define the gradient and divergence operations. The symbol ∇ is refered to as *nabla* and, when applied to a scalar field, is known as the gradient and sometimes denoted by $\mathrm{grad} := \nabla$. Likewise the divergence, defined below, is sometimes written $\mathrm{div} := \nabla_\bullet$.

- Given scalar field $\varphi : \mathbb{R}^d \to \mathbb{R}$, we define vector field $\nabla\varphi : \mathbb{R}^d \to \mathbb{R}^d$, the *gradient* of φ, by:

$$\left\{\nabla\varphi(x)\right\}_i = \frac{\partial\varphi}{\partial x_i}(x) = \varphi_{,i}(x) \tag{3.1}$$

- Given vector field $\psi : \mathbb{R}^d \to \mathbb{R}^d$, we define scalar field $\nabla_\bullet\psi : \mathbb{R}^d \to \mathbb{R}$, the *divergence* of ψ, by:

$$\nabla_\bullet\psi(x) = \sum_{i=1}^{d} \frac{\partial\psi_i}{\partial x_i}(x) = \psi_{i,i}(x) \tag{3.2}$$

Here we employ the index notation in which repeated indices denote summation and index i after a comma denotes differentiation with respect to x_i.

Example 3.1. Given scalar field $\varphi : \mathbb{R}^d \to \mathbb{R}$, define scalar field $\Delta\varphi : \mathbb{R}^d \to \mathbb{R}$ (called the *Laplacian* of φ) by

$$\Delta\varphi = \nabla_\bullet\left(\nabla\varphi\right) = \sum_{i=1}^{d} \frac{\partial}{\partial x_i}\left(\frac{\partial\varphi}{\partial x_i}\right) = \sum_{i=1}^{d} \frac{\partial^2\varphi}{\partial x_i^2} = \varphi_{,ii}.$$

◊

Example 3.2. Consider the scalar fields $p : \mathbb{R}^d \to \mathbb{R}$ and $a : \mathbb{R}^d \to \mathbb{R}$. Then the vector field ∇p may be multiplied pointwise by scalar field a to obtain vector field $a\nabla p$. Hence the expression $\nabla_\bullet \underbrace{\left(a\nabla p\right)}_{\text{vector field}}$ defines a scalar field. ◊

3.2 Link to Graph Laplacian

Throughout this section, we will use following notation: $x^{(j)} \in D \subset \mathbb{R}^d$, D is a bounded, open set in $\mathbb{R}^d$, and $j \in \mathsf{Z} := \{1, \ldots, N\}$. This is a special case of Lecture 1 in which we have specialized from data in a general metric space to data that lies in a bounded subset of $\mathbb{R}^d$; we will use the Euclidean norm $\|\cdot\|$ to define the metric. Furthermore, we assume that the $x^{(j)}$ are instances of i.i.d. random variables drawn from a probability measure on D with density ρ; that is, $x^{(j)} \sim \rho$, where ρ is a scalar field satisfying $\int_D \rho(x)\,\mathrm{d}x = 1$ and $\rho(x) \geqslant 0$ for all $x \in D$. We study the limit

of the graph Laplacian when $N \to \infty$ and when a particular scaling of the weight function η is made.

Recall that the graph Laplacian leads to the quadratic form

$$\langle \varphi, L\varphi \rangle_{\mathbb{R}^N} = \frac{1}{2} \sum_{(i,j)} W_{ij} |\varphi_i - \varphi_j|^2.$$

We can think of φ as a function $\varphi : \mathsf{Z} \to \mathbb{R}$, or, equivalently, $\varphi \in \mathbb{R}^N$. In our setting, every point in Z is associated to a point in D. It is thus natural to shift our perspective to consider $\varphi : D \to \mathbb{R}$, or, equivalently, to view φ as a scalar field. Another way to think about this transition is to imagine there is a function that maps each input to an output, and now we will be working with the function itself: φ such that $\varphi_i = \varphi\left(x^{(i)}\right)$.

We also assume that the weight function η is parameterized by ε and write it as $\eta_\varepsilon : \mathbb{R}^+ \to \mathbb{R}^+$; typically, this scaling is $\eta_\varepsilon(r) \coloneqq \varepsilon^{-d} \eta(r/\varepsilon)$. The weights are then defined as

$$W_{ij} = \eta_\varepsilon \left(\left\| x^{(i)} - x^{(j)} \right\| \right).$$

We assume that $\varepsilon \ll 1$ is chosen so that

$$\int_{\mathbb{R}^+} \eta_\varepsilon(r) \psi(r) \, \mathrm{d}r \approx \psi(0),$$

that is, η_ε is approximately a Dirac delta function.

Making use of all the assumptions, we can write:

$$\begin{aligned}
\langle \varphi, L\varphi \rangle_{\mathbb{R}^N} &= \frac{1}{2} \sum_{i,j} \eta_\varepsilon \left(\left\| x^{(i)} - x^{(j)} \right\| \right) \left| \varphi\left(x^{(i)}\right) - \varphi\left(x^{(j)}\right) \right|^2 \\
&\propto \frac{1}{N^2} \sum_{i,j} \eta_\varepsilon \left(\left\| x^{(i)} - x^{(j)} \right\| \right) \left| \frac{\varphi\left(x^{(i)}\right) - \varphi\left(x^{(i)}\right)}{\varepsilon} \right|^2 \\
&\approx \int_D \int_D \eta_\varepsilon \left(\|x - y\| \right) \left| \frac{\varphi(x) - \varphi(y)}{\varepsilon} \right|^2 \rho(x) \rho(y) \, \mathrm{d}x \mathrm{d}y \qquad (3.3a) \\
&\approx \int_D \left| \nabla \varphi(x) \right|^2 \rho(x)^2 \, \mathrm{d}x \qquad (3.3b) \\
&= \int_D \langle \nabla \varphi(x), \left(\rho(x)^2 \nabla \varphi(x) \right) \rangle_{\mathbb{R}^d} \, \mathrm{d}x \qquad (3.3c) \\
&= - \int_D \varphi(x) \nabla_\bullet \left(\rho(x)^2 \nabla \varphi(x) \right) \, \mathrm{d}x. \qquad (3.3d)
\end{aligned}$$

Here, approximate equality (3.3a) is obtained by letting $N \to \infty$ and making use of the *law of large numbers*; approximate equality (3.3b) is

achieved when $\varepsilon \to 0$ and employing the fact that η_ε is an approximation of the Dirac delta function; finally, equation (3.3d) is a simple application of Green's identity,[1] provided that $\nabla\varphi(x)_\bullet n = 0$ on the boundary ∂D, where n is the outward pointing unit normal on ∂D. Also note, in equations (3.3b)–(3.3c), notation $|\cdot|$, applied to an element of $\mathbb{R}^d$, simply denotes Euclidean norm on $\mathbb{R}^d$: $|\cdot| = \|\cdot\|$ and $\langle\cdot,\cdot\rangle_{\mathbb{R}^d}$ the corresponding inner product.

Rewriting (3.3d) demonstrates the approximation

$$\langle \varphi, L\varphi\rangle_{\mathbb{R}^N} \approx C\langle \varphi, \mathcal{L}\varphi\rangle_{L^2_\rho} \tag{3.4}$$

where on the right hand side we have transitioned into viewing φ as a function defined on the domain D. In equation (3.4), we introduced the following notation: that of the inner product on function space L^2_ρ and of the operator $\mathcal{L}$:

$$\langle a, b\rangle_{L^2_\rho} = \int_D a(x)b(x)\rho(x)\,\mathrm{d}x,$$
$$\mathcal{L}\varphi = \underbrace{-\frac{1}{\rho}}_{\text{scalar field.}}\nabla_\bullet \underbrace{\left(\rho^2\nabla\varphi\right)}_{\text{vector field.}}, \quad \nabla\varphi_\bullet n = 0 \text{ on } \partial D.$$

The norm on L^2_ρ is

$$\|a\|_{L^2_\rho} = \left(\int_D \rho(x)|a(x)|^2 dx\right)^{\frac{1}{2}}.$$

With these concepts introduced we see that $\mathcal{L}\varphi\colon \mathbb{R}^d \to \mathbb{R}$; hence $\mathcal{L}$ maps scalar fields into scalar fields and this operator may be thought of as the limit of the graph Laplacian. The operator $\mathcal{L}$ is in fact coincident (up to a multiplicative factor) with the Laplacian defined in Example 3.1 if ρ is constant. The eigenvalues and eigenvectors of this operator provide understanding of data clustering via the graph Laplacian, in the large data limit $N \gg 1$; these eigenvalues and eigenvectors are dependent on the data distribution ρ which, in turn, reflects clustering present in the data. This discussion serves to motivate eigenvalue problems in PDEs.

[1] This is simply integration by parts in dimension $d = 1$.

3.3 Flow in Porous Media

This section presents another example of how differential operators, similar to those in the last section, arise in applications, this time drawn from physics. Modeling flow in porous media arises, for instance, in design of reservoir dams; the model we describe is a convenient and physically reasonable way to describe macroscopic behavior of flow in porous media without describing the pore structure at a miscroscale.

We introduce the following notation. We take $D \subset \mathbb{R}^d$ (where here $d = 2$ or 3 for problems of physical interest, and $d = 1$ is of interest for illustrative purposes and for computational examples) and define the scalar fields

$$\begin{aligned} a &: D \to \mathbb{R}, \quad \text{permeability} \\ f &: D \to \mathbb{R}, \quad \text{sources/sinks of fluid} \\ p &: D \to \mathbb{R}, \quad \text{pressure.} \end{aligned}$$

and the vector field $v : D \to \mathbb{R}^d$, given by $v = -a\nabla p$; this is the velocity field, and the relationship between permeability, pressure and velocity is known as *Darcy's law*. Conservation of mass imposes the condition that $\nabla_\bullet v = f$. Thus, given a and f and solving for p, the flow can then be modeled by the following PDE:

$$\begin{cases} -\nabla_\bullet \left(a\nabla p\right) = f & \text{in } D, \\ p = 0 & \text{on } \partial D. \end{cases} \tag{PDE}$$

We imposed the condition that the pressure is zero on the boundary $\partial D = \overline{D}\backslash D$; other boundary conditions may be more natural in applications, but this *homogeneous Dirichlet boundary condition* provides the simplest setting in which to describe the mathematical structure.

In order to understand the motivation for our analysis of (PDE), we refer to a finite-dimensional analogy, encapsulated in the following lemma.

Lemma 3.3. *For a fixed matrix $A \in \mathbb{R}^{N\times N}$ and vector $f \in \mathbb{R}^N$, equality $Ap = f$ holds if and only if the following is true:*

$$\langle Ap, q\rangle_{\mathbb{R}^N} = \langle f, q\rangle_{\mathbb{R}^N} \quad \forall q \in \mathbb{R}^N.$$

A similar idea may be used for (PDE): we take the inner product of the equation, in a suitable inner product space, with arbitrary function q. Unlike the finite-dimensional case this does not lead to an equivalent problem to (PDE), but rather to a weakened form (wPDE). Solutions

of (wPDE) coincide with those of (PDE) if there is enough smoothness in the functions a, f and hence in the solution p. We take the L^2 inner product with q in (PDE) to obtain:

$$\langle -\nabla_{\bullet}\left(a\nabla p\right), q\rangle_{L^2} = \langle f, q\rangle_{L^2} \quad \forall q \in V, \tag{3.5}$$

where the space L^2 is L^2_ρ with $\rho \equiv 1$, and V is the following space:

$$V \coloneqq H^1_{0,a}(D) = \left\{ u : D \to \mathbb{R} \,\middle|\, u(\partial D) = 0,\ \|u\|_{H^1_{0,a}} < \infty \right\}.$$

The inner product and norm on $H^1_{0,a}$ are defined as follows:

$$\begin{aligned} \langle u, v\rangle_{H^1_{0,a}} &= \int_D a(x)\langle \nabla u(x), \nabla v(x)\rangle_{\mathbb{R}^d}\,\mathrm{d}x, \\ \|u\|^2_{H^1_{0,a}} &= \langle u, u\rangle_{H^1_{0,a}}. \end{aligned}$$

Using Green's theorem (integrating by parts in dimensions $d \geq 1$) in (3.5), we write the problem (PDE) in weak form:

$$\text{Find } p \in V: \qquad \langle p, q\rangle_{H^1_{0,a}} = \langle f, q\rangle_{L^2} \quad \forall q \in V. \tag{wPDE}$$

Remark 3.4. *Provided $f \in L^2$, this definition makes sense because the fact that $q \in V$ implies $q \in L^2$. In fact, because $q \in V$ (which is much smaller than L^2), the space in which f can be chosen may be enlarged to a set V^* (the dual space of V) which contains L^2 as a strict subset.*

3.4 Galerkin Approximation of (wPDE)

Let $\varphi_j \in V$, $j \in \{1, \ldots, J\}$; we introduce a finite-dimensional space V^J which will allow us to formulate a finite-dimensional analog of (wPDE):

$$V^J \coloneqq \left\{ v \in V \,\middle|\, \exists \left\{v_j\right\}_{j=1}^J,\ v_j \in \mathbb{R},\ \text{with } v = \sum_{j=1}^J v_j \varphi_j \right\}.$$

The approximation of (wPDE) is then formulated as follows:

$$\text{Find } p^J \in V^J: \qquad \langle p^J, q\rangle_{H^1_{0,a}} = \langle f, q\rangle_{L^2} \quad \forall q \in V^J. \tag{wPDE^J}$$

The next theorem shows that p^J is the best possible approximation to p from the space V^J, in a well-defined sense:

Theorem 3.5.

$$\|p^J - p\|_{H^1_{0,a}} \leqslant \|q^J - p\|_{H^1_{0,a}} \quad \forall q^J \in V^J.$$

Proof

$$\begin{aligned}\|q^J - p\|^2_{H^1_{0,a}} &= \|p^J - p + q^J - p^J\|^2_{H^1_{0,a}} \\ &= \|p^J - p\|^2_{H^1_{0,a}} + 2\underbrace{\langle p^J - p, q^J - p^J\rangle_{H^1_{0,a}}}_{=0 \quad (+)} + \|q^J - p^J\|^2_{H^1_{0,a}}.\end{aligned}$$

Therefore, assuming that (+) holds,

$$\|q^J - p\|^2_{H^1_{0,a}} = \|p^J - p\|^2_{H^1_{0,a}} + \|q^J - p^J\|^2_{H^1_{0,a}} \geqslant \|p^J - p\|^2_{H^1_{0,a}}.$$

It only remains to show (+). Since (wPDE) and (wPDE^J) hold $\forall q \in V^J$, we have:

$$\begin{aligned}&\langle p, q\rangle_{H^1_{0,a}} = \langle p^J, q\rangle_{H^1_{0,a}} && \forall q \in V^J \\ &\langle p - p^J, q\rangle_{H^1_{0,a}} = 0 && \forall q \in V^J.\end{aligned}$$

Since $q^J - p^J \in V^J$, the last identity implies that

$$\langle p - p^J, q^J - p^J\rangle_{H^1_{0,a}} = 0. \tag{+}$$

□

Remark 3.6. *To solve* (wPDE^J)*, write* $p^J = \sum_{j=1}^J u_j\varphi_j$ *and note that solving it for any* $q \in V^J$ *is equivalent to solving the following:*

$$\left\langle \sum_{j=1}^J u_j\varphi_j, \varphi_k \right\rangle_{H^1_{0,a}} = \left\langle f, \varphi_k \right\rangle_{L^2}, \quad k = 1, \ldots, J.$$

Thus, we simply need to solve $Au = b$*, where* A*,* u *and* b *are defined in the following way:*

$$u = \begin{pmatrix} u_1 \\ \vdots \\ u_J \end{pmatrix}, \quad A = \begin{pmatrix} a_{11} & \ldots & a_{1J} \\ \vdots & & \vdots \\ a_{J1} & \ldots & a_{JJ} \end{pmatrix}, \quad b = \begin{pmatrix} b_1 \\ \vdots \\ b_J \end{pmatrix} \tag{3.6}$$
$$a_{jk} = \langle \varphi_j, \varphi_k\rangle_{H^1_{0,a}}, \quad b_k = \langle f, \varphi_k\rangle_{L^2}.$$

3.5 What Ideas Did We Need?

(i) Probability spaces (Lecture 4).
(ii) L^p and H^1 spaces (Lectures 5 and 6).
(iii) H^1_0 space (Lecture 8).

(iv) Orthogonality (Lecture 11).
(v) Lax-Milgram Theorem (Lecture 12).

Exercises

3.1 Let $a, b\colon \mathbb{R}^d \to \mathbb{R}$ be scalar fields, and $\psi\colon \mathbb{R}^d \to \mathbb{R}^d$ be a vector field. Prove the following identities:

(a) $\nabla(ab) = a\nabla b + b\nabla a$,
(b) $\nabla_\bullet(a\psi) = a\nabla_\bullet\psi + \psi_\bullet\nabla a$,
(c) $a\Delta b - b\Delta a = \nabla_\bullet(a\nabla b - b\nabla a)$.

3.2 Consider the function $\eta : \mathbb{R}^d \to \mathbb{R}^+$ defined by

$$\eta(x) \coloneqq \begin{cases} C\exp\left(\dfrac{1}{|x|^2-1}\right), & \text{if } |x| < 1, \\ 0, & \text{otherwise,} \end{cases}$$

where C is defined such that $\int_{\mathbb{R}^d} \eta(x)\mathrm{d}x = 1$ and $|\cdot|$ denotes Euclidean norm on $\mathbb{R}^d$. This function is called the "standard mollifier" and plays an important role in theory of distributions (generalized functions). Further, we define $\eta_\varepsilon(x) \coloneqq \frac{1}{\varepsilon^d}\eta(\frac{x}{\varepsilon})$.

Show that:

(a) $\int_{\mathbb{R}^d} \eta_\varepsilon(x)\mathrm{d}x = 1$;
(b) both $\eta(x)$ and $\eta_\varepsilon(x)$ are C^∞ functions, i.e., any k-th partial derivative exists for any natural k;
(c) **(hard)** the *mollification* $f_\varepsilon \coloneqq f * \eta_\varepsilon$ of f defined by

$$f_\varepsilon(x) = \int_D \eta_\varepsilon(x-y)f(y)\mathrm{d}y$$

is also a C^∞ function, for any $f\colon D \to \mathbb{R}$ such that $\int_D f(y)\mathrm{d}y$ exists, where $D \subset \mathbb{R}^d$ is open and bounded.

3.3 Suppose we wish to solve (wPDE$^{\mathrm{J}}$) in the one-dimensional case, i.e., $d = 1$, and, for simplicity, $D = (s, r)$. How does the support of the φ_j's affect the sparsity of matrix A in (3.6)? Hence, what would a computationally beneficial choice of φ_j's be?

4

Topological, Metric, Probability and Vector Spaces

In this lecture we study various spaces of elements and highlight the relationships between the resulting spaces. In doing so, we set the course itself, which is primarily concerned with vector spaces, in a broader context. We introduce the central concepts of normed vector spaces and inner product spaces.

4.1 Topological and Metric Spaces

Definition 4.1. *Let X be a set, and let $\mathcal{U}$ be a collection of subsets of X. $\mathcal{U}$ is called a* topology *on X if*

(i) $X, \varnothing \in \mathcal{U}$;
(ii) $\{U_\alpha\}_\alpha$ *is an arbitrary collection of elements of* $\mathcal{U}$ *implies that* $\bigcup_\alpha U_\alpha \in \mathcal{U}$;
(iii) $\{U_j\}_{j=1}^n$ *is a finite collection of elements of* $\mathcal{U}$ *implies that* $\bigcap_{j=1}^n U_j \in \mathcal{U}$.

The pair $(X, \mathcal{U})$ is called a topological space.

Example 4.2. Let $X = \mathbb{R}$. We define $\mathcal{U}$ as follows: we say that $U \in \mathcal{U}$ if and only if it can be represented as a union of open intervals, that is, $U = \bigcup_\alpha (l_\alpha, r_\alpha)$. Then $(X, \mathcal{U})$ is an example of a topological space. ◊

Definition 4.3. *Let X be a set. A function $d\colon X \times X \to \mathbb{R}$ is said to be a* metric *on X if*

(i) $d(u, v) \geq 0$ *for all* $u, v \in X$ *and* $d(u, v) = 0$ *if and only if* $u = v$;
(ii) $d(u, v) = d(v, u)$ *for all* $u, v \in X$;
(iii) $d(u, v) \leq d(u, w) + d(w, v)$ *for all* $u, v, w \in X$.

The pair (X, d) is called a metric space.

Example 4.4. Given any set X, define $d(x,x) = 0$ and $d(x,y) = 1$ if $y \neq x$. This is known as the *discrete metric* on X because every point is distance 1 from every other point. ◊

Definition 4.5. *A subset S of a metric space (X,d) is called open if, given any point $x \in S$, there exists real $\delta > 0$ such that given any point $y \in X$ with $d(x,y) < \delta$, it follows that $y \in S$.*

Definition 4.6. *A* Borel set *is any set in a metric space that can be formed from open sets (or, equivalently, from closed sets) through the operations of countable union, countable intersection, and complement.*

The notion of open sets, and hence of Borel set, extends to topological spaces, but for simplicity we have chosen not to discuss open sets for arbitrary topological spaces, only for metric spaces. This is because, in the metric space setting, open sets have a more intuitive and direct connection to open sets in $\mathbb{R}^d$.

Proposition 4.7. *From every metric space it is possible to create a topological space.*

Proof Let (X,d) be a metric space and define

$$B(a,r) = \{x \in X : d(a,x) < r\} \quad \forall a \in X,\ r \in \mathbb{R}^+, \tag{4.1}$$

noting that $r = 0$ gives $\emptyset$. Then define

$$\mathcal{U} = \bigcup_{\alpha} B(a_\alpha, r_\alpha) \quad \text{over all } a_\alpha \in X,\ r_\alpha \in \mathbb{R}^+.$$

□

We refer to $B(a,r)$ as the open ball in X, of radius r, centered at a. Notice that the construction of the topology in the previous example is precisely the construction used in Example 4.2.

Definition 4.8. *A topological space $(X,\mathcal{U})$ is said to be* metrizable *if there exists a metric on X that generates the topology $\mathcal{U}$ by means of the construction used in the proof of Proposition 4.7.*

Remark 4.9. *Not every topological space is metrizable.*

Definition 4.10. *Let (X,d_X), (Y,d_Y) be two metric spaces. The map $f : X \to Y$ is* continuous at point $v \in X$ *if for any $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that*

$$d_X(v,x) < \delta$$

implies that

$$d_Y(f(v), f(x)) < \varepsilon$$

for all $x \in X$.

Example 4.11. Let $\mathsf{Z} \coloneqq \{1, \ldots, N\}$ denote the nodes of a graph and A the adjacency matrix of the graph so that $A_{ij} = 1$ if $(i, j) \in \mathsf{E}$, the edge set, and $A_{ij} = 0$ if $(i, j) \notin \mathsf{E}$. A path in the graph node set Z is the ordered set

$$\gamma = \{w_0, w_1, \cdots, w_n\}$$

with $A_{w_m, w_{m+1}} = 1$ for $m = 0, \ldots, n-1$. The path connects u and v if $w_0 = u$ and $w_n = v$. The length of γ, denoted $\ell(\gamma)$, is n. Then, extending the codomain of a metric to $\mathbb{R} \cup \{+\infty\}$, we can define it as

$$d(u, v) = \begin{cases} \min\{\ell(\gamma) : \gamma \text{ connects } u, v\}, & \text{if } \exists \text{ at least one such } \gamma, \\ +\infty, & \text{otherwise.} \end{cases}$$

Below we introduce vector spaces, and norms, demonstrating a setting in which there is a canonical way to introduce a metric, using the norm. This graph-based example is interesting because the metric is defined on a space that is not a vector space. ◊

4.2 Measurable Spaces and Probability Spaces

Definition 4.12. *Given a set* X*, a* σ-algebra over X *is a collection of subsets* Σ *of* X *with the properties that*

- $X \in \Sigma$;
- Σ *is closed under complements;*
- Σ *is closed under countable unions.*

Example 4.13. The *Borel* σ*-algebra* is a canonical σ-algebra associated with a metric space X; it is the collection of all Borel sets on X. The Borel σ-algebra on X may also be characterized as the smallest σ-algebra containing all open sets (or, equivalently, all closed sets) in X. ◊

Because the notion of open sets also has a meaning in arbitrary topological spaces, the notion of Borel σ-algebra extends to topological spaces. However, recall that we have chosen not to discuss open sets at this level of generality.

Definition 4.14. *Let X be a set and Σ a σ-algebra over X. The pair (X, Σ) is referred to as a* measurable space. *A function $\mu : \Sigma \to \mathbb{R} \cup \{-\infty, +\infty\}$ is a* measure *if it satisfies the following properties:*

- *non-negativity: for all $E \in \Sigma$, we have $\mu(E) \geq 0$;*
- *null empty set: $\mu(\emptyset) = 0$;*
- *countable additivity (or σ-additivity): for all countable collections $\{E_i\}_{i=1}^{\infty}$ of pairwise disjoint sets in Σ,*

$$\mu\left(\bigcup_{k=1}^{\infty} E_k\right) = \sum_{k=1}^{\infty} \mu(E_k).$$

The triplet (X, Σ, μ) is referred to as a measure space. *If, in addition, $\mu : \Sigma \to [0, 1]$ and $\mu(X) = 1$, then it is a* probability space *and μ is a* probability measure.

The notion of *signed measure* extends measure to allow negative values on some elements of the σ-algebra.

Example 4.15. Consider the measurable space $(\mathbb{R}, \Sigma)$ with Σ being the Borel σ-algebra. Probability measure is a map $\mu : \Sigma \to [0, 1]$ such that $\mu(\mathbb{R}) = 1$, $\mu(\emptyset) = 0$. We can then define

$$\mathbb{E}^{\mu} f = \int_{\mathbb{R}} f(x)\mu(\mathrm{d}x), \text{ with}$$
$$\mu(A) = \int_A \mu(dx) = \mathbb{E}^{\mu} \mathbb{1}_A.$$

Here $\mathbb{1}_A(x) = 1$ if $x \in A$ and is 0 otherwise; it is the *indicator function* of the set A. Let $\mathcal{M}_+$ be the collection of all such probability measures. Define the distance between two measures

$$d(\mu, \nu) = \sup_{f \in \mathsf{F}} |\mathbb{E}^{\mu} f - \mathbb{E}^{\nu} f|,$$

where

$$\mathsf{F} = \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; \sup_{x \in \mathbb{R}} |f(x)| \leq 1 \right\}.$$

The metric $d : \mathcal{M}_+ \times \mathcal{M}_+ \to \mathbb{R}$ is the *total variation metric*. ◊

4.3 Vector Spaces

Definition 4.16. *Let $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. A* vector space *V over $\mathbb{K}$ is a collection of elements (vectors), and operations $+, \times$ satisfying:*

- $u + v \in V$ *for all* $u, v \in V$; $\alpha \times v \in V$ *for all* $\alpha \in \mathbb{K}$, $v \in V$;
- *associativity:* $u + (v + w) = (u + v) + w$ *for all* $u, v, w \in V$;
- *commutativity:* $u + v = v + u$ *for all* $u, v \in V$;
- *identity +: there exists* $0 \in V$ *such that* $v + 0 = v$ *for all* $v \in V$;
- *inverse +: for all* $v \in V$ *there exists* $-v \in V$ *such that* $v + (-v) = 0$;
- *compatibility:* $\lambda(\mu v) = (\lambda\mu)v$ *for all* $\lambda, \mu \in \mathbb{K}$ *and* $v \in V$;
- *identity* $\times$: $1v = v$ *for all* $v \in V$;
- *distributivity in* V: $\lambda(u + v) = \lambda u + \lambda v$ *for all* $\lambda \in \mathbb{K}$ *and* $u, v \in V$;
- *distributivity in* $\mathbb{K}$: $(\lambda + \mu)v = \lambda v + \mu v$ *for all* $\lambda, \mu \in \mathbb{K}$ *and* $v \in V$.

Example 4.17. The following four examples, and variants on them, will be primary examples of vector spaces in these lectures:

- $V = \mathbb{R}^n$ is a vector space over $\mathbb{K} = \mathbb{R}$.
- $V = \mathbb{C}^{m \times n}$ is a vector space over $\mathbb{K} = \mathbb{C}$.
- $V = \{f : \mathbb{Z}^+ \to \mathbb{R}\}$ is a vector space over $\mathbb{K} = \mathbb{R}$.
- $V = \{f : \mathbb{R}^d \to \mathbb{R}\}$ is a vector space over $\mathbb{K} = \mathbb{R}$.

$\Diamond$

Remark 4.18. *The first example may be reformulated in the same format as the third and fourth by defining* $\mathsf{Z} := \{1, \ldots, n\}$ *and noting that*

$$\mathbb{R}^n = \{f : \mathsf{Z} \to \mathbb{R}\}.$$

Variants on the second example that we will consider include replacing $\mathbb{C}^{m \times n}$ *and* $\mathbb{C}$ *by* $\mathbb{R}^{m \times n}$ *and* $\mathbb{R}$*, as well as linear operators between more general vector spaces. A variant on the third example that we will consider includes infinite sequences of vectors in* $\mathbb{R}^q$ *rather than just in* $\mathbb{R}$*. A variant on the fourth example that we will consider is the setting where the domain of* f *is* D *a bounded open subset of* $\mathbb{R}^d$*.*

Definition 4.19. *A subset* U *in a vector space* V *is* convex *if, for all* $u, v \in U$ *and all* $\lambda \in [0, 1]$*, the element* $\lambda u + (1 - \lambda)v$ *is also in* U*.*

Definition 4.20. *A* subspace *is a vector space* V' *that is a subset of some larger vector space* V*.*

Sometimes the term *linear subspace* is used to distinguish from other types of subspace, but we will only use linear subspaces here and hence dispense with the word linear. The terminology *vector subspace* is also used in some texts.

Example 4.21. The set of probability measures on $\mathbb{R}$, $\mathcal{M}_+$ defined in Example 4.15, is not a vector space over $\mathbb{R}$, (with + and $\times$ defined in the natural way). Indeed, let $\mu, \nu \in \mathcal{M}_+$ noting that $\pi(\mathbb{R}) = 1$ for any $\pi \in \mathcal{M}_+$. Then, $a\mu + b\nu \notin \mathcal{M}_+$ unless $a + b = 1$. To see this note that

$$(a\mu + b\nu)(\mathbb{R}) = a\mu(\mathbb{R}) + b\nu(\mathbb{R}) = a + b$$

Thus, if $a\mu + b\nu \in \mathcal{M}_+$, then $a + b = 1$. ◊

4.4 Normed Vector Spaces

Definition 4.22. *A* norm *on vector space V is a map $\|\cdot\| : V \to \mathbb{R}$ satisfying*

(i) $\|x\| \geq 0$ *for all $x \in V$, with $\|x\| = 0$ if and only if $x = 0$;*
(ii) $\|\alpha x\| = |\alpha|\|x\|$ *for all $\alpha \in \mathbb{K}$, $x \in V$;*
(iii) $\|x + y\| \leq \|x\| + \|y\|$ *for all $x, y \in V$.*

Remark 4.23. *A norm is a continuous function from V into $\mathbb{R}$. To see this, note that*

$$\|a + h\| \leq \|a\| + \|h\|, \quad \|a\| \leq \|a + h\| + \|h\|$$

implying that

$$\big|\|a + h\| - \|a\|\big| \leq \|h\|.$$

Continuity follows.

Definition 4.24. *Let $(V, \|\cdot\|)$ be a normed vector space. We define the* open ball *of radius $r > 0$ at $v \in V$ by*

$$B_V(v, r) = \{u \in V : \|u - v\| < r\}.$$

This is coincides with the definition (4.1) in the metric space setting if $d(a, b) := \|a - b\|$. This is a conseqeunce of the general fact which now follows.

Proposition 4.25. *Every normed vector space is a metric space.*

Proof Define $d(u, v) = \|u - v\|$. Clearly,

(i) $d(u, v) = \|u - v\| \geq 0$ for all $u, v \in V$ and $d(u, v) = 0$ if and only if $u - v = 0$;
(ii) $d(u, v) = d(v, u) = \|u - v\|$;
(iii) $d(u, v) = \|u - v\| = \|u - w + w - v\| \leq \|u - w\| + \|w - v\| = d(u, w) + d(w, v)$.

□

Examples 4.26. In all the following examples of normed vector spaces, we take the field $\mathbb{K}$ to be $\mathbb{R}$; extension to the complex setting is straightforward.

- $\mathbb{R}^n$ with $\|u\|_p = \left(\sum_{j=1}^{n} |u_j|^p\right)^{1/p}$, $1 \le p < \infty$.
- $\mathbb{R}^n$ with $\|u\|_\infty = \max_j |u_j|$.
- The space $\ell^p(\mathbb{N}; \mathbb{R}^m)$ of all functions $f : \mathbb{N} \to \mathbb{R}^m$, such that their ℓ^p-norm is finite, $1 \le p < \infty$:

$$\|u\|_{\ell^p} = \left(\sum_{j=1}^{\infty} \|u_j\|_{\mathbb{R}^m}^p\right)^{1/p}, \tag{4.2}$$

  where $\|\cdot\|_{\mathbb{R}^m}$ is any norm on $\mathbb{R}^m$. In the case $m = 1$ this may also be viewed as a subset of $\mathbb{R}^\infty$ comprising elements for which the norm in (4.2), with $\|\cdot\|_{\mathbb{R}^m}$ replaced by the real absolute value $|\cdot|$, is finite.
- $\mathbb{R}^\infty$ with $\|u\|_{\ell^\infty} = \sup_{j \in \mathbb{Z}^+} |u_j| < \infty$.
- The space $L^p(D; \mathbb{R}^m)$ of all functions $f : D \to \mathbb{R}^m$, $D \subset \mathbb{R}^d$ bounded and open, such that their L^p-norm is finite, $1 \le p < \infty$:

$$\|u\|_{L^p} = \left(\int_D \|u(x)\|_{\mathbb{R}^m}^p dx\right)^{1/p},$$

  where $\|\cdot\|_{\mathbb{R}^m}$ is any norm on $\mathbb{R}^m$.

◊

Remark 4.27. *The functions in ℓ^p with domain $\mathbb{N}$ are easily extended to other domains, such as $\mathbb{Z}^+$. Likewise the functions in L^p with domain D may be extended to domain $D = \mathbb{R}^d$.*

Examples 4.28. Let $A \in \mathbb{R}^{n \times n}$; any of the following defines a norm, and hence, induces a normed vector space:

- maximum norm: $\|A\|_{\max} := \max_{i,j} |A_{ij}|$
- Frobenius norm: $\|A\|_{\mathrm{F}} := \sqrt{\sum_{i,j=1}^{n} |A_{ij}|^2} = \sqrt{\mathrm{tr}\,(A^\top A)}$

- p-norm (induced by a $\mathbb{R}^n$ p-norm): $\|A\|_p \coloneqq \sup_{\|x\|_p=1} \|Ax\|_p$

$\Diamond$

Definition 4.29. *Two norms $\|\cdot\|_a$, $\|\cdot\|_b$ on a vector space V are* equivalent *if there exist constants $0 < c_1 \leq c_2 < \infty$ such that, for all $x \in V$,*

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

Theorem 4.30. *All norms on $\mathbb{R}^n$ are* equivalent*: for every two norms $\|\cdot\|_{\mathrm{a}}$ and $\|\cdot\|_{\mathrm{b}}$ on $\mathbb{R}^n$, there exist constants $0 < c_1 \leq c_2 < \infty$ such that, for all $x \in \mathbb{R}^n$,*

$$c_1\|x\|_{\mathrm{a}} \leq \|x\|_{\mathrm{b}} \leq c_2\|x\|_{\mathrm{a}}.$$

Proof Using property (ii) from the definition of a vector norm it suffices to consider vectors $x \in S \coloneqq \left\{x \in \mathbb{R}^n \;\middle|\; \|x\|_2 = 1\right\}$. Since $\|\cdot\|_{\mathrm{a}}$ is non-zero on all of S, the function $f\colon S \to \mathbb{R}$ given by $f(x) \coloneqq \|x\|_{\mathrm{b}}/\|x\|_{\mathrm{a}}$ is well-defined. The function f is continuous when extended to $\mathbb{R}^n\backslash\{0\}$, by Remark 4.23. Because the set S is a closed, bounded subset of $\mathbb{R}^n$, and because f is continuous on S, there exist $x_1, x_2 \in S$ with $f(x_1) \leq f(x) \leq f(x_2)$ for all $x \in S$. Setting $c_1 \coloneqq f(x_1) > 0$ and $c_2 \coloneqq f(x_2) < \infty$ completes the proof. □

Example 4.31. The norm on $\ell^p(\mathbb{Z}^+;\mathbb{R}^m)$ requires a norm on $\mathbb{R}^m$. However, two equivalent norms on $\mathbb{R}^m$ lead to two equivalent norms on $\ell^p(\mathbb{Z}^+;\mathbb{R}^m)$. To see this, let $\|\cdot\|_a$, $\|\cdot\|_b$ be two equivalent norms on $\mathbb{R}^m$ and $\{u_j\}_{j=1}^{\infty} \in \ell^p$:

$$\begin{aligned} c_1\|u\|_a \leq \|u\|_b \leq c_2\|u\|_a \quad \forall u \in \mathbb{R}^m, \text{ hence} \\ c_1^p\|u_j\|_a^p \leq \|u_j\|_b^p \leq c_2^p\|u_j\|_a^p \quad \forall j \in \mathbb{Z}^+. \end{aligned}$$

Summing the first n of such inequalities, we obtain:

$$c_1^p \sum_{j=0}^{n} \|u_j\|_a^p \leq \sum_{j=0}^{n} \|u_j\|_b^p \leq c_2^p \sum_{j=0}^{n} \|u_j\|_a^p.$$

Since $u \in \ell_a^p$ and $u \in \ell_b^p$, which implies that both $\|u\|_{\ell_a^p}$ and $\|u\|_{\ell_b^p}$ exist, it only remains to take the limit $n \to \infty$ to obtain the needed fact:

$$\begin{aligned} c_1^p\|u\|_{\ell_a^p}^p \leq \|u\|_{\ell_b^p}^p \leq c_2^p\|u\|_{\ell_a^p}^p, \text{ or} \\ c_1\|u\|_{\ell_a^p} \leq \|u\|_{\ell_b^p} \leq c_2\|u\|_{\ell_a^p}. \end{aligned}$$

$\Diamond$

Example 4.32. Consider $\mathbb{R}^d$ equipped with a norm, noting that this is also a metric space and so we can define the open ball $B(a, r)$ for any $a \in \mathbb{R}^d$ and $r > 0$ using Definition 4.24. This is a convex set. For $d > 1$ and the Euclidean norm, the union of two balls $B(a_i, r_i)$, $i = 1, 2$, is convex if and only if one is a subset of the other. ◊

4.5 Inner Product Spaces

Definition 4.33. *An* inner product *on a vector space V is a function $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{K}$ satisfying the following four conditions:*

(i) $\langle v, v \rangle \geq 0$ *for all* $v \in V$, *with* $\langle v, v \rangle = 0$ *if and only if* $v = 0$;
(ii) $\langle v, w \rangle = \overline{\langle w, v \rangle}$ *for all* $v, w \in V$;
(iii) $\langle v, \alpha w \rangle = \alpha \langle v, w \rangle$ *for all* $\alpha \in \mathbb{K}$, $v, w \in V$;
(iv) $\langle u, v + w \rangle = \langle u, v \rangle + \langle u, w \rangle$ *for all* $u, v, w \in V$.

Proposition 4.34. *Every inner product space is a normed vector space with norm* $\|u\| = (\langle u, u \rangle)^{1/2}$.

Lemma 4.35 (Cauchy–Schwarz)**.** *Let $\langle \cdot, \cdot \rangle \colon V \times V \to \mathbb{K}$ be an inner product. Then*

$$\left| \langle x, y \rangle \right| \leq \|x\| \, \|y\| \tag{4.3}$$

for every $x, y \in V$, and equality holds if and only if x and y are linearly dependent.

Proof If either x or y is zero then the result holds with equality and trivial linear dependence. Hence we assume for the remainder of the proof that both x and y are non-zero. For every $\lambda \in \mathbb{K}$ we have

$$0 \leq \langle x - \lambda y, x - \lambda y \rangle = \langle x, x \rangle - \overline{\lambda} \langle y, x \rangle - \lambda \langle x, y \rangle + \lambda \overline{\lambda} \langle y, y \rangle. \tag{4.4}$$

For $\lambda = \langle x, y \rangle / \langle y, y \rangle$, this becomes

$$0 \leq \langle x, x \rangle - \frac{\langle x, y \rangle \langle y, x \rangle}{\langle y, y \rangle} - \frac{\langle y, x \rangle \langle x, y \rangle}{\langle y, y \rangle} + \frac{\langle y, x \rangle \langle x, y \rangle}{\langle y, y \rangle} = \|x\|^2 - \frac{\left| \langle x, y \rangle \right|^2}{\|y\|^2}.$$

Multiplying the result by $\|y\|^2$ gives the desired inequality.

If equality holds in (4.3), then $x - \lambda y$ in (4.4) must be 0 and thus x and y are linearly dependent. If, on the other hand, x and y are linearly dependent, say $x = \alpha y$ for some $\alpha \in \mathbb{K}$, then $\lambda = \langle y, \alpha y \rangle / \langle y, y \rangle = \alpha$ and $x - \lambda y = 0$ giving equality in (4.4) and thus in (4.3). □

Proof of Proposition 4.34.
In order to prove that an inner product space is a normed vector space, we only need to show that the function defined by $\|\cdot\| = (\langle\cdot,\cdot\rangle)^{1/2}$ is indeed a norm, that is, satisfies the three norm axioms. The first axiom follows directly from the first axiom of inner product; the second one follows, respectively, from axioms (ii) and (iii). Thus the triangle inequality is the only non-trivial item we are left to prove.

Note that, using the Cauchy–Schwarz inequality,

$$\begin{aligned}\|x+y\|^2 &= \langle x+y, x+y\rangle\\ &= \langle x,x\rangle + \langle x,y\rangle + \langle y,x\rangle + \langle y,y\rangle\\ &= \|x\|^2 + 2\,\mathrm{Re}\big(\langle x,y\rangle\big) + \|y\|^2\\ &\le \|x\|^2 + 2\big|\langle x,y\rangle\big| + \|y\|^2\\ &\le \|x\|^2 + 2\|x\|\|y\| + \|y\|^2\\ &= \big(\|x\| + \|y\|\big)^2.\end{aligned}$$

Taking the square root on both sides completes the proof. □

Examples 4.36.

- $\mathbb{R}^n$ over $\mathbb{R}$ with $\langle u,v\rangle = \sum_{j=1}^{n} u_j v_j$ is called *Euclidean* space.
- $\ell^2(\mathbb{N};\mathbb{R})$ with $\langle u,v\rangle = \sum_{j=1}^{\infty} u_j v_j$.
- $\ell^2(\mathbb{N};\mathbb{R}^m)$ with $\langle u,v\rangle = \sum_{j=1}^{\infty} \langle u_j, v_j\rangle_{\mathbb{R}^m}$.
- $L^2(D;\mathbb{R})$ with
$$\langle u,v\rangle = \int_D u(x)v(x)\,dx. \tag{4.5}$$
- $L^2(D;\mathbb{R}^m)$ with
$$\langle u,v\rangle = \int_D \langle u(x), v(x)\rangle_{\mathbb{R}^m} dx. \tag{4.6}$$

See Exercises 6.4, 6.10. ◊

Definition 4.37. *Let V be an inner product space. Given $a, b \in V$, we define the* outer product $a \otimes b$ *by*

$$(a\otimes b)c = \langle b,c\rangle a$$

for any $c \in V$.

Example 4.38. The outer product of two vectors $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$ is the matrix $a \otimes b \in \mathbb{R}^{m \times n}$ with entries $(a \otimes b)_{ij} = a_i b_j$. In this finite-dimensional setting it is commonplace to write $a \otimes b = ab^\top$. ◊

Exercises

4.1 Show that the following relations hold for all $x \in \mathbb{C}^n$:

(a) $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$
(b) $\|x\|_\infty \leq \|x\|_2 \leq \sqrt{n}\|x\|_\infty$
(c) $\|x\|_\infty \leq \|x\|_1 \leq n\|x\|_\infty$

4.2 Prove that, if $A \in \mathbb{C}^{n \times n}$ is an invertible matrix and $\|\cdot\|_{(1)}$ is a vector norm on $\mathbb{C}^n$, then $\|\cdot\|_{(2)}$ defined by $\|x\|_{(2)} := \|Ax\|_{(1)}$ is also a vector norm on $\mathbb{C}^n$.

4.3 Prove that, for Hermitian positive definite $A \in \mathbb{C}^{n \times n}$, the relation

$$\langle x, y \rangle_A := \langle x, Ay \rangle$$

for all $x, y \in \mathbb{C}^n$ defines an inner product on $\mathbb{C}^n$. Show also that $\|x\|_A := (\langle x, Ax \rangle)^{\frac{1}{2}}$ defines a norm. If $\|\cdot\|$ is any norm on $\mathbb{C}^{n \times n}$, then show that $\|A \cdot\|$ also defines a norm on $\mathbb{C}^{n \times n}$.

4.4 Prove that for all matrices $A \in \mathbb{C}^{m \times n}$,

$$\|A\|_{\max} \leq \|A\|_{\mathrm{F}} \leq \sqrt{mn}\|A\|_{\max}.$$

4.5 For $A, B \in \mathbb{C}^{n \times n}$, let $\langle A, B \rangle = \mathrm{tr}(A^* B)$, where tr denotes the trace of a matrix.

(a) Show that $\langle \cdot, \cdot \rangle$ is an inner product on $\mathbb{C}^{n \times n}$.
(b) Show that $\|A\|_{\mathrm{F}}^2 = \langle A, A \rangle$ for all $A \in \mathbb{C}^{n \times n}$.

4.6 Show that $\|\cdot\|_{\max}$ is a vector norm on the space of $n \times n$ matrices, but not an induced matrix norm.

4.7 Prove that the 2-norm of a matrix is unchanged by multiplication by unitary matrices, i.e., for any $A \in \mathbb{R}^{m \times n}$, unitary $U \in \mathbb{R}^{m \times m}$ and unitary $V \in \mathbb{R}^{n \times n}$,

$$\|UAV\|_2 = \|A\|_2.$$

4.8 Prove that the Frobenius norm of a matrix is unchanged by multiplication by unitary matrices. This is an analogue of Exercise 4.7.

4.9 Prove that the sphere $S(a, r) := \{v \in \mathbb{R}^n : \|v - a\| < r\}$ is convex. What can you say about the convexity of the union of two spheres $S(a, r)$ and $S(b, s)$?

4.10 Give an example of a linear space X and a metric $d(\cdot, \cdot)$ on X to show that the function $f(u) = d(u, 0)$ does not necessarily define a norm on X.

4.11 In Example 4.21 we show that the set of probability measures on $\mathbb{R}$ is not a vector space over $\mathbb{R}$, with the natural definitions of $+$ and $\times$. Is the same true for the set of measures on $\mathbb{R}$? What about the set of signed measures on $\mathbb{R}$?

4.12 Show that the discrete metric in Example 4.4 is indeed a metric.

4.13 Show that d from Example 4.11 is indeed a metric.

4.14 Let $X = \mathbb{R}$ and let $F : \mathbb{R} \to \mathbb{R}$ be a monotonic increasing function. Show that $d(x, y) = |F(x) - F(y)|$ is a metric. Where in the proof do you use that F is monotonic increasing? Could you use F monotonic decreasing?

5
Banach and Hilbert Spaces

In the previous lecture we demonstrated the nesting of topological spaces, metric spaces, normed vector spaces and inner product spaces. In this lecture we add the additional structure of completeness. This leads to the notion of Banach spaces and Hilbert spaces which are complete normed vector spaces and inner product spaces respectively – see Figure 5.1. Completeness may also be discussed in the context of metric spaces, although we do not pursue this here; see Figure 5.2. We start by discussing the concepts of open and closed sets and then move on to define completeness and hence Banach and Hilbert space. We then consider ℓ^p and L^p as canonical examples of Banach and (for $p = 2$) Hilbert spaces, establishing this in the former case.

5.1 Open and Closed

In a normed vector space $(V, \|\cdot\|)$, we can discuss the concept of limits of sequences $\{v_n\}$ and hence of a closed set. This will enable us to define, in the following section, the concept of Cauchy sequences and hence completeness.

Definition 5.1. *Let $(V, \|\cdot\|)$ be a normed vector space. A sequence $\{v_n\}$ in V* converges *to $v \in V$ if, for any $\epsilon > 0$, there is $N = N(\epsilon) \in \mathbb{N}$ such that, for all $n > N$,*

$$\|v_n - v\| < \epsilon.$$

We say that v is the limit point *of convergent sequence $\{v_n\}$ in V.*

Definition 5.2. *A* closed set *S in V is one which contains all limit points of*

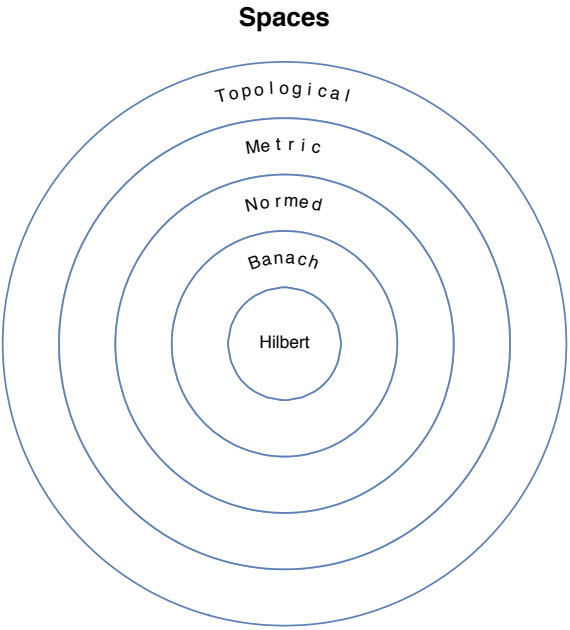


Figure 5.1 An illustration of how the different classes of spaces introduced are contained within one another.

sequences in S. If the set S in V is not closed, then the closure *of S, denoted $\overline{S}$, comprises S and the limit points of all convergent sequences in S.*

Example 5.3. Let $V = \mathbb{R}$ and use modulus as norm; this defines a normed vector space. Consider the sequence $\{v_n\}$ defined by $v_n = n^{-1}$. Then 0 is the limit of convergent sequence $\{v_n\}$. ◊

Definition 5.4. *Let $(V, \|\cdot\|)$ be a normed vector space. An* open set *S in V is one which, for every $v \in S$, there is $\delta > 0$ such that $B_V(v, \delta) \subset S$.*

Example 5.5. For $w \in V$, $S = B_V(w, r)$ is open. If $v \in S$ then there is $\delta \in (0, r)$ such that $\|v - w\| \leq r - \delta$. Thus $S' = B_V(v, \delta/2) \subset S$ because if $u \in S'$ then

$$\|u - w\| = \|(u - v) + (v - w)\| \leq \|u - v\| + \|v - w\| < \frac{1}{2}\delta + r - \delta < r.$$

Furthermore,

$$\overline{B_V(w, r)} = \{u \in V : \|u - w\| \leq r\}.$$

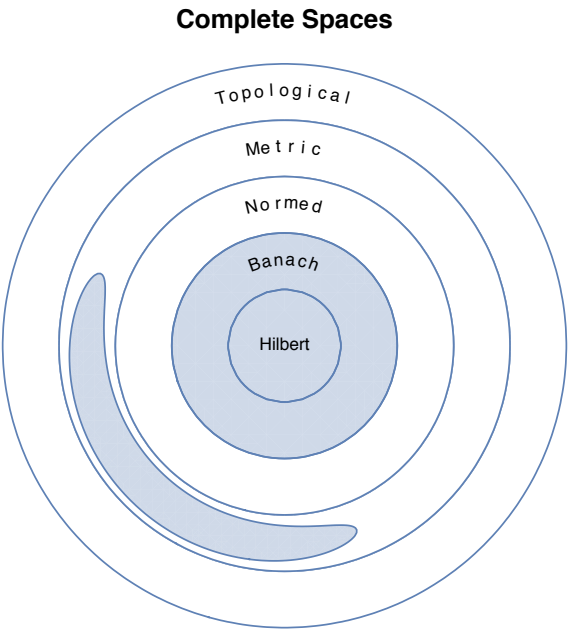


Figure 5.2 Complete spaces as a subset of the classes of spaces introduced. This figure re-enforces the fact that completeness may be defined in a metric space, although we define it only in a normed vector space.

To see this let $\{s^{(n)}\}$ denote a sequence in S with limit point v. Then, for any $\epsilon > 0$ there is $N = N(\epsilon)$ such that, for $n \geq N$,

$$\|v - w\| = \|(s^{(n)} - w) + (v - s^{(n)})\| < r + \epsilon.$$

Since ϵ is arbitrary it follows that $\|v - w\| \leq r$. The proof of closure is finalized by showing that any point in $\partial S = \overline{B_V(w, r)} \backslash B_V(w, r)$ can be attained as limit of sequence $s^{(n)} = w + u^{(n)}$ with $u^{(n)}$ converging to u with property $u - w$ in ∂S. ◊

5.2 Completeness

Definition 5.6. *A sequence $\{v_n\}_{n=1}^{\infty}$ in a normed vector space $(V, \|\cdot\|)$ is* Cauchy *if, for every $\epsilon > 0$, there exists $N > 0$ such that $\|v_n - v_m\| < \epsilon$ for all $n, m \geq N$. A normed vector space V is* complete *if every Cauchy sequence converges: there is an element $v \in V$ such that $\|v_n - v\| \to 0$ as $n \to \infty$.*

Example 5.7. $\mathbb{R}^n$ with any norm $\|\cdot\|$ is complete. Indeed all finite-dimensional vector spaces are complete. ◊

Example 5.8. Returning to Example 5.3, we note that, for $n \geq m$,

$$|v_n - v_m| \leq m^{-1}.$$

By choosing $N > \epsilon^{-1}$ we see that the sequence is Cauchy. The preceding example tells us that it is then necessarily convergent. ◊

Example 5.9. Let $P = \{f : [0,1] \to \mathbb{R},\ f \text{ polynomial}\}$, the collection of all real polynomials on the unit interval $[0,1]$. We make this a normed vector space by defining

$$\|f\|_P := \sup_{x\in[0,1]} |f(x)|.$$

Now consider the polynomial sequence $\{p^{(n)}\}$ defined by

$$p^{(n)}(x) = \sum_{j=1}^{n} \frac{x^j}{j!}.$$

Then $\{p^{(n)}\}$ is Cauchy in $(P, \|\cdot\|_P)$. However, any limit of the sequence $p^{(n)}(\cdot)$ must equal $\exp(\cdot)$ since uniform convergence implies pointwise convergence and, pointwise in $x \in [0,1]$, $p^{(n)}(x) \to \exp(x)$ as a real-valued sequence. But $\exp(\cdot) \notin P$. ◊

Definition 5.10. *A complete normed vector space is called a* Banach space.

Example 5.11. The vector space $\mathbb{R}^n$ equipped with any norm is a Banach space. We will use this fact below in establishing that the ℓ^p spaces are Banach spaces. ◊

Remark 5.12. *There is a way of* completing *any normed vector space. Often the completion is intuitive (for example, by adding continuous functions that can be uniformly approximated by polynomials as in Example 5.9). However, the general resolution of this issue is rather subtle and beyond the scope of these notes; a related issue is briefly discussed in Section 8.4.*

Definition 5.13. *An inner product space $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ which is complete as a normed vector space is a* Hilbert space.

Example 5.14. $\mathbb{R}^n$ with Euclidean inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|_2$ is a Hilbert space. ◊

5.3 Some Key Inequalities

Definition 5.15. *We say that two indices $p, q : 1 \le p, q \le \infty$, are* conjugate *if $p^{-1} + q^{-1} = 1$.*

Lemma 5.16 (Young's inequality)**.** *Let $a, b > 0$ and let (p, q) be conjugate indices with $1 < p, q < \infty$. Then*

$$ab \le \frac{1}{p}a^p + \frac{1}{q}b^q.$$

Proof The function $\exp(\cdot)$ is convex: its second derivative is non-negative. As a consequence, for any $\lambda \in [0, 1]$,

$$\exp\big(\lambda x + (1-\lambda)y\big) \le \lambda \exp(x) + (1-\lambda)\exp(y)$$

and so

$$\begin{aligned} ab &= \exp(\log a + \log b) = \exp\Big(\frac{1}{p}\log a^p + \frac{1}{q}\log b^q\Big) \\ &\le \frac{1}{p}\exp\big(\log(a^p)\big) + \frac{1}{q}\exp\big(\log(b^q)\big) = \frac{1}{p}a^p + \frac{1}{q}b^q. \end{aligned}$$

□

Lemma 5.17 (Hölder inequality)**.** *Let $v \in \ell^p$ and $w \in \ell^q$ with p, q conjugate and $1 \le p, q \le \infty$. Then $z := (v_1 w_1, v_2 w_2, \cdots) \in \ell^1$ and*

$$\|z\|_{\ell^1} \le \|v\|_{\ell^p}\|w\|_{\ell^q}. \tag{5.1}$$

Proof We start with the case $1 < p, q < \infty$. By the Young's inequality we have

$$\sum_{j=1}^{n} \frac{|v_j||w_j|}{\|v\|_{\ell^p}\|w\|_{\ell^q}} \le \sum_{j=1}^{n}\Big(\frac{1}{p}\frac{|v_j|^p}{\|v\|_{\ell^p}^p} + \frac{1}{q}\frac{|w_j|^q}{\|w\|_{\ell^q}^q}\Big) \le \frac{1}{p} + \frac{1}{q} = 1.$$

Thus for each n we have

$$\sum_{j=1}^{n} |v_j||w_j| \le \|v\|_{\ell^p}\|w\|_{\ell^q}$$

and (5.1) follows. Now consider the case $p = 1$ and $q = \infty$ (noting that $p = \infty$ and $q = 1$ is then covered by symmetry). For every $n \in \mathbb{N}$, we have

$$\sum_{j=1}^{n} |v_j||w_j| \le \max_{1\le j\le n} |w_j| \sum_{j=1}^{n} |v_j| \le \|v\|_{\ell^1}\|w\|_{\ell^\infty}.$$

□

Remark 5.18. *The following identities, equivalent when $p, q \in [1, \infty]$ and hence useful when discussing conjugacy, hold:*

$$\frac{1}{p} + \frac{1}{q} = 1 \iff p + q = pq \iff p = (p-1)q.$$

Lemma 5.19 (Minkowski inequality)**.** *If $1 \le p \le \infty$ and $v, w \in \ell^p$, then $v + w \in \ell^p$ and*

$$\|v + w\|_{\ell^p} \le \|v\|_{\ell^p} + \|w\|_{\ell^p}.$$

Proof First consider the case $p \in (1, \infty)$, with q the conjugate exponent. Notice that

$$\left|\frac{1}{2}(a+b)\right|^p \le \frac{1}{2}(|a|^p + |b|^p)$$

since the mapping $x \mapsto |x|^p$ is convex on $\mathbb{R}$. From this it follows that

$$2^{-(p-1)}|v_j + w_j|^p \le (|v_j|^p + |w_j|^p).$$

Thus it is clear that $v + w \in \ell^p$. Now by the Hölder inequality we have, for each $n \in \mathbb{N}$,

$$\begin{aligned}
\sum_{j=1}^{n} |v_j + w_j|^p &\le \sum_{j=1}^{n} |v_j + w_j|^{p-1}|v_j| + \sum_{j=1}^{n} |v_j + w_j|^{p-1}|w_j| \\
&\le \Big(\sum_{j=1}^{n} |v_j + w_j|^{(p-1)q}\Big)^{1/q} \Big(\big(\sum_{j=1}^{n} |v_j|^p\big)^{1/p} + \big(\sum_{j=1}^{n} |w_j|^p\big)^{1/p} \Big) \\
&\le \Big(\sum_{j=1}^{n} |v_j + w_j|^p\Big)^{1/q} \big(\|v\|_{\ell^p} + \|w\|_{\ell^p}\big)
\end{aligned}$$

so that the desired result follows:

$$\begin{aligned}
\|v + w\|_{\ell^p} &= \lim_{n\to\infty} \Big(\sum_{j=1}^{n} |v_j + w_j|^p\Big)^{1/p} \\
&= \lim_{n\to\infty} \Big(\sum_{j=1}^{n} |v_j + w_j|^p\Big)^{1-1/q} \\
&\le \|v\|_{\ell^p} + \|w\|_{\ell^p}.
\end{aligned}$$

The cases $p = 1$ and $p = \infty$ are left as an exercise. □

5.4 ℓ^p and L^p are Banach Spaces

Theorem 5.20. *For every $1 \le p \le \infty$, the sequence space ℓ^p is a Banach space.*

Proof The norm axioms are straightforward to verify for the definition of $\|\cdot\|_{\ell^p}$, noting that the Minkowski inequality gives the triangle property. Thus it remains to show that the normed vector space is complete, i.e., that every Cauchy sequence is convergent to a limit in ℓ^p. Let $\{v^{(k)}\}$ denote a Cauchy sequence in ℓ^p so that, for every $\epsilon > 0$, there is $N = N(\epsilon)$ such that for every $n, m \ge N$,

$$\|v^{(n)} - v^{(m)}\|_{\ell^p}^p = \sum_{j=1}^{\infty} |v_j^{(n)} - v_j^{(m)}|^p < \epsilon^p. \tag{5.2}$$

This shows, in particular, that for any fixed j, the sequence $\{v_j^{(n)}\}$ in $\mathbb{R}$ is Cauchy and hence (since $\mathbb{R}$ is a Banach space with any norm) has limit v_j'. We define the candidate limit $v' = (v_1', v_2', \dots)$. We want to show that $v' \in \ell^p$ and that $\|v^{(n)} - v'\|_{\ell^p} \to 0$. From (5.2) we have

$$\sum_{j=1}^{M} |v_j^{(n)} - v_j^{(m)}|^p \le \sum_{j=1}^{\infty} |v_j^{(n)} - v_j^{(m)}|^p < \epsilon^p.$$

Letting $m \to \infty$ we obtain

$$\sum_{j=1}^{M} |v_j^{(n)} - v_j'|^p < \epsilon^p$$

and since this holds for all M we have

$$\sum_{j=1}^{\infty} |v_j^{(n)} - v_j'|^p \le \epsilon^p$$

establishing that $v^{(n)} - v' \in \ell^p$. Since ℓ^p is a vector space and $v^{(n)} \in \ell^p$, we deduce from the Minkowski inequality that $v' \in \ell^p$ and that, for any $\epsilon > 0$ and $n \ge N(\epsilon)$, $\|v^{(n)} - v'\|_{\ell^p} \le \epsilon$ establishing convergence of $v^{(n)}$ to v'. The case $p = \infty$ is left as an exercise. □

Theorem 5.21. *Let $\langle \cdot, \cdot \rangle_{\mathbb{R}^n}$ be the standard Euclidean inner product. Define for $a, b \in \ell^2(\mathbb{Z}^+; \mathbb{R}^n)$*

$$\langle a, b \rangle_{\ell^2} = \sum_{j=1}^{\infty} \langle a_j, b_j \rangle_{\mathbb{R}^n}.$$

Then $\ell^2(\mathbb{Z}^+; \mathbb{R}^n)$ is a Hilbert space with this inner product and resulting norm.

Example 5.22. Let $v_j = j^{-s}$, $s \geq 0$. If $s = 0$, $v \in \ell^\infty$, but $v \notin \ell^p$ for $p \in [1, \infty)$. If $s > 0$,

$$\|v\|_{\ell^p}^p = \sum_{j=1}^{\infty} j^{-sp} < \infty$$

if and only if $p > \frac{1}{s}$. ◊

Remark 5.23. *This demonstrates that the norms $\|\cdot\|_{\ell^a}$, $\|\cdot\|_{\ell^b}$ are not equivalent unless $a = b$.*

Theorem 5.24. *Let $u : D \to \mathbb{R}^n$ and define*

$$\|u\|_{L^p}^p = \int_D \|u(x)\|^p \mathrm{d}x,$$

with $\|\cdot\|$ being any norm on $\mathbb{R}^n$. For every $1 \leq p \leq \infty$, the function space $L^p(D;\mathbb{R}^n) = \{u : D \to \mathbb{R}^n \mid \|u\|_{L^p} < \infty\}$ is a Banach space.

Theorem 5.25. *Let $\langle\cdot,\cdot\rangle_{\mathbb{R}^n}$ be the standard Euclidean inner product. Define for $a, b \in L^2(D;\mathbb{R}^n)$*

$$\langle a, b\rangle_{L^2} \coloneqq \int_D \langle a(x), b(x)\rangle_{\mathbb{R}^n} \,\mathrm{d}x. \tag{5.3}$$

Then $L^2(D;\mathbb{R}^n)$ is a Hilbert space with this inner product and resulting norm.

Exercise 6.4 establishes that expression (5.3) does indeed define an inner product for the case $n = 1$, while exercise 6.10 introduces Banach spaces $L^p(D;\mathbb{R}^n)$ for $p \in [1, \infty]$.

Example 5.26. Let $D = B(0,1) = \{x \in \mathbb{R}^d : |x| < 1\}$. Let $v(x) = |x|^{-s}$, $s \geq 0$. If $s = 0$, then $v \in L^p(D;\mathbb{R})$ for $p \in [1, \infty]$. If $s > 0$, then

$$\|v\|_{L^p}^p = \int_D |x|^{-sp} \mathrm{d}x \propto \int_0^1 r^{-sp} r^{d-1} \mathrm{d}r.$$

which is integrable if $1 - d + sp < 1$, i.e., $p < d/s$. Thus $v \in L^p(D;\mathbb{R})$ if and only if $p \in [1, d/s)$. ◊

Remark 5.27. *This demonstrates that the norms $\|\cdot\|_{L^a}$, $\|\cdot\|_{L^b}$ are not equivalent unless $a = b$.*

Exercises

5.1 Let (M, d) be a metric space, and let $\{u_n\}_{n\in\mathbb{N}}$ be a Cauchy sequence on M. Show that if $\{u_n\}_{n\in\mathbb{N}}$ has a convergent subsequence, then $\{u_n\}_{n\in\mathbb{N}}$ is convergent.

5.2 Let $\ell_0 \subseteq \mathbb{R}^\infty$ be the set of sequences with finitely many non-zero entries. Show that ℓ_0 is not complete with respect to the ℓ^2-norm.

5.3 Let $\alpha > 0$ and define the sequences $v, w \in \mathbb{R}^\infty$ by

$$v_j = \frac{1}{j^{2+\alpha}}, \quad w_j = j^\alpha.$$

Show that $v \in \ell^2$, $w \notin \ell^2$, but

$$\sum_{j=1}^{\infty} v_j w_j < \infty.$$

If both $v, w \notin \ell^2$, is it possible for the sum to converge? What if we insist all terms are non-negative?

5.4 Let $(X, \|\cdot\|)$ be a Banach space. Prove that $u \mapsto \|u\|^2$ is continuous. If X is a real Hilbert space can any more be said about the regularity of this map?

5.5 *(Generalized Hölder)* Let $r \in [1, \infty)$ and $p_1, \dots, p_k \in [1, \infty]$ be such that

$$\sum_{j=1}^{k} \frac{1}{p_j} = \frac{1}{r}.$$

Show that if $v_j \in \ell^{p_j}$ for each j, then $v_1 \cdots v_k \in \ell^r$ and

$$\|v_1 \cdots v_k\|_{\ell^r} \leq \|v_1\|_{\ell^{p_1}} \|v_2\|_{\ell^{p_2}} \cdots \|v_k\|_{\ell^{p_k}}$$

where multiplication is performed component-wise.

5.6 Let X, Y be Banach spaces. Show that a map $f : X \to Y$ is continuous if and only if $f^{-1}(V) \subseteq X$ is open for any open $V \subseteq Y$.

5.7 Find an example of a sequence $\{u_n\}$ on a Banach space X such that $u_n \rightharpoonup u$ for some $u \in X$ and

$$\liminf_{n\to\infty} \|u_n\| > \|u\|.$$

5.8 Let $(X, \|\cdot\|)$ be finite dimensional. Show that weak convergence in X is equivalent to strong convergence in X.

5.9 Prove that if a subsequence of a Cauchy sequence in a normed vector space converges, then the entire sequence converges.

5.10 Prove that if a sequence converges, then the limit is unique.

5.11 Let $C(I;\mathbb{R})$ denote continuous real-valued functions on the unit interval $I = [0,1]$. Show that this forms a vector space over $\mathbb{R}$. Consider the family of functions $\{f^{(n)}\}_{n\in\mathbb{N}}$ defined by

$$f^{(n)}(x) = \left\{ \begin{array}{c} 0, \quad x \in [0, \frac{1}{2} - \frac{1}{n}) \\ 1 - n(\frac{1}{2} - x), \quad x \in [\frac{1}{2} - \frac{1}{n}, \frac{1}{2}] \\ 1, \quad x \in x \in (\frac{1}{2}, 1]. \end{array} \right\}.$$

Using these, show that $C(I;\mathbb{R})$ is not complete with respect to the $L^1(I;\mathbb{R})$ norm.

6
Spaces of Functions and Fourier Analysis

In the previous lecture we introduced Banach spaces and Hilbert spaces, which are complete normed vector spaces and inner product spaces, respectively. We illustrated these spaces with spaces of sequences, ℓ^p, and spaces of functions, L^p. In this lecture, we look further at examples of Banach and Hilbert spaces of functions and links to Fourier analysis. In particular, we introduce spaces of continuous functions and, using the notion of weak derivatives, of Sobolev spaces of functions. The notion of Fourier transform also serves as an example of a linear transformation between Banach or Hilbert spaces and serves as motivation for Chapter 7.

6.1 Spaces of Continuous Functions

Here we introduce several concepts relating to continuity. The resulting spaces of continuous functions are themselves interesting examples of Banach spaces; they are also foundational in the definition of Sobolev spaces in the next section.

Definition 6.1. *An n-dimensional* multi-index *is a vector $\alpha \in (\mathbb{Z}^+)^n$, $\alpha = (\alpha_1, \ldots, \alpha_n)$. Furthermore, its* order *$|\alpha|$ is defined as*

$$|\alpha| = \alpha_1 + \alpha_2 + \cdots + \alpha_n.$$

Let $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be a vector and $\partial = (\partial_1, \partial_2, \cdots, \partial_n)$ be a vector of differential operators (we abbreviate $\partial/\partial x_j$ to ∂_j). With this notation, we define

$$x^\alpha = x_1^{\alpha_1} \cdots x_n^{\alpha_n}, \qquad \partial^\alpha = \partial_1^{\alpha_1} \cdots \partial_n^{\alpha_n} = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_n^{\alpha_n}}.$$

Example 6.2. Suppose that $u : \mathbb{R}^2 \to \mathbb{R}$, that is, $(x_1, x_2) \mapsto u(x_1, x_2) \in \mathbb{R}$. Then

$$\sum_{|\alpha|=2} \partial^\alpha u = \frac{\partial^2 u}{\partial x_1^2} + 2\frac{\partial^2 u}{\partial x_1 \partial x_2} + \frac{\partial^2 u}{\partial x_2^2},$$

and Leibniz's rule becomes

$$\partial^\alpha (uv) = \sum_{\beta \le \alpha} \binom{\alpha}{\beta} \partial^{\alpha-\beta} u \, \partial^\beta v,$$

where

$$\binom{\alpha}{\beta} = \frac{\alpha!}{(\alpha-\beta)!\beta!},$$

$\alpha! = \alpha_1! \cdots \alpha_n!$, and $\beta \le \alpha$ if $\beta_i \le \alpha_i$ for all $i = 1, \ldots, n$. ◊

Let $D \subseteq \mathbb{R}^d$ be an open subset of $\mathbb{R}^d$.

Definition 6.3. *The space $C(\overline{D}; \mathbb{R})$ comprises real-valued functions which are continuous on the closure of D, $\overline{D}$; some sources also use the notation C^0.*

Theorem 6.4. *The space $C(\overline{D}; \mathbb{R})$ is a Banach space when equipped with the norm*

$$\|u\|_C := \sup_{x \in D} |u(x)|.$$

Definition 6.5. *The space $C^k(\overline{D}; \mathbb{R})$ comprises all real-valued functions for which each derivative up to order $k \in \mathbb{Z}^+$ exists and is continuous on $\overline{D}$:*

$$C^k(\overline{D}; \mathbb{R}) := \left\{ u : \overline{D} \to \mathbb{R} \;\middle|\; \partial^\alpha u \in C(\overline{D}; \mathbb{R}),\ |\alpha| \le k \right\}.$$

Theorem 6.6. *The space $C^k(\overline{D}; \mathbb{R})$ is a Banach space when equipped with the norm*

$$\|u\|_{C^k} := \sum_{0 \le |\alpha| \le k} \sup_{x \in D} |(\partial^\alpha u)(x)|.$$

Definition 6.7. *The space $C^\infty(\overline{D}; \mathbb{R})$ consists of all infinitely differentiable real-valued functions:*

$$C^\infty(\overline{D}; \mathbb{R}) := \left\{ u : \overline{D} \to \mathbb{R} \;\middle|\; \partial^\alpha u \in C(\overline{D}; \mathbb{R})\ \forall \alpha \in (\mathbb{Z}^+)^d \right\} = \bigcap_{k=0}^{\infty} C^k(\overline{D}; \mathbb{R})$$

Definition 6.8. *For any function $u : D \to \mathbb{R}$, we define the* support *of u, denoted as* $\operatorname{supp} u$*, as the smallest closed set containing the set*

$$\left\{ x \in D : u(x) \neq 0 \right\}.$$

Definition 6.9. *The space $C_c^\infty(\overline{D};\mathbb{R})$ (pronounced as "C infinity compact") consists of all infinitely differentiable real-valued functions whose support is a compact set:*

$$C_c^\infty(\overline{D};\mathbb{R}) := \left\{ u : D \to \mathbb{R} \;\middle|\; u \in C^\infty(\overline{D};\mathbb{R}) \text{ and } \operatorname{supp} u \text{ is compact} \right\}.$$

This space is also often called the *space of test functions* due to its role in definitions of other spaces. Intuitively, functions and operators that cannot be defined in a straightforward manner are *tested* against C_c^∞ functions; thus, they can be defined by their *action* on test functions.

Remark 6.10. *Everything in this section is readily generalized to functions on D taking values in $\mathbb{R}^m$, rather than $\mathbb{R}$; we leave the details to the reader.*

6.2 Sobolev Spaces

This section is devoted to a particular family of Banach and Hilbert spaces of functions, called Sobolev spaces. These spaces are particularly useful for the study of differential equations and signal and image processing; but, these are by no means the only contexts in which they appear. Our point of departure is the key concept of *weak derivative*.

For the rest of this section, we will restrict D to be a bounded open subset of $\mathbb{R}^d$; however, similar notions can be introduced for the unbounded case. Now suppose that u is differentiable on D. Then for any test function $\varphi \in C_c^\infty(\overline{D};\mathbb{R})$, one can integrate by parts to obtain

$$\int_D \frac{\partial u}{\partial x_j}\,\varphi \,\mathrm{d}x = -\int_D u\,\frac{\partial \varphi}{\partial x_j}\,\mathrm{d}x, \tag{6.1}$$

using the fact that φ has compact support on D (so the boundary term vanishes). Repeating this process $|\alpha|$ times yields

$$\int_D (\partial^\alpha u)\,\varphi \,\mathrm{d}x = (-1)^{|\alpha|}\int_D u\,\partial^\alpha \varphi \,\mathrm{d}x$$

for any multi-index α.

Now, while the left-hand side of (6.1) makes sense only if $\partial_j u$ exists, the right-hand side makes sense for any $u \in L^1$ since we always have $\partial_j \varphi \in L^\infty$.

Definition 6.11. *For a function $u \in L^1(D;\mathbb{R})$, we say that v is the* weak

derivative *of u with respect to x_j, and write $\partial_j u = v$, if $v \in L^1(D;\mathbb{R})$ and*

$$\int_D v\,\varphi\,\mathrm{d}x = -\int_D u\,\partial_j\varphi\,\mathrm{d}x$$

for every $\varphi \in C_c^\infty(\overline{D};\mathbb{R})$. In a similar fashion we say that v is the the weak derivative of order *α of u and write $\partial^\alpha u = v$, if $v \in L^1(D;\mathbb{R})$ and*

$$\int_D v\,\varphi\,\mathrm{d}x = (-1)^{|\alpha|}\int_D u\,\partial^\alpha\varphi\,\mathrm{d}x$$

for every $\varphi \in C_c^\infty(\overline{D};\mathbb{R})$.

The following proposition is fundamental.

Proposition 6.12. *When they exist, weak derivatives are unique. Furthermore, they coincide with the classical notion of derivative when the classical (i.e., strong) derivative exists.*

Definition 6.13. *The* Sobolev space $W^{k,p}(D;\mathbb{R})$ *is defined as*

$$W^{k,p}(D;\mathbb{R}) = \left\{u : D \to \mathbb{R} \;\middle|\; \partial^\alpha u \in L^p(D;\mathbb{R}),\ 0 \le |\alpha| \le k\right\},$$

with norm

$$\|u\|_{W^{k,p}} = \left(\sum_{0\le|\alpha|\le k}\|\partial^\alpha u\|_{L^p}^p\right)^{1/p},$$

where derivatives $\partial^\alpha u$ are understood in the weak sense.

In other words, the Sobolev space $W^{k,p}(D;\mathbb{R})$ is a space of functions whose weak derivatives up to order k exist and belong to $L^p(D;\mathbb{R})$ (including the function itself since $k \in \mathbb{Z}^+$, so in particular, $k = 0$).

Theorem 6.14. *The space $W^{k,p}(D;\mathbb{R})$ is a Banach space for $1 \le p \le \infty$.*

Definition 6.15. *The* Sobolev space $H^k(D;\mathbb{R})$ *is defined to be $W^{k,2}(D;\mathbb{R})$.*

Theorem 6.16. *The space $H^k(D;\mathbb{R})$ is a Hilbert space when equipped with the inner product*

$$\langle u, v\rangle_{H^k} = \sum_{0\le|\alpha|\le k}\langle\partial^\alpha u, \partial^\alpha v\rangle_{L^2},$$

and the norm

$$\|u\|_{H^k} = \left(\sum_{0\le|\alpha|\le k}\|\partial^\alpha u\|_{L^2}^2\right)^{1/2}.$$

Example 6.17. Recall the vector field $\nabla u : D \to \mathbb{R}^d$ defined by

$$\nabla u = (\partial_1 u, \dots, \partial_d u)^\top.$$

Then

$$\|u\|_{H^1}^2 = \langle u, u \rangle_{H^1} = \|u\|_{L^2}^2 + \sum_{j=1}^{d} \|\partial_j u\|_{L^2}^2 = \|u\|_{L^2}^2 + \|\nabla u\|_{L^2}^2.$$

◊

Remark 6.18. *Of particular interest in certain elliptic partial differential equations is the Sobolev space $H_0^1(D;\mathbb{R})$, which is defined in Lecture 8.*

Remark 6.19. *Everything in this section is readily generalized to functions on D taking values in $\mathbb{R}^m$, rather than $\mathbb{R}$; we leave the details to the reader.*

6.3 Fourier Transform

Definition 6.20. *The* Schwarz space *is defined by*

$$S(D;\mathbb{C}) := \left\{ u \in C^\infty(\overline{D};\mathbb{C}) \;\middle|\; \sup_{x \in D} |x^\alpha \left(\partial^\beta u\right)(x)| < \infty \quad \forall \alpha, \beta \in (\mathbb{Z}^+)^d \right\}.$$

Definition 6.21. *If $u \in S(\mathbb{R}^d;\mathbb{C})$, we define the* Fourier transform *by*

$$\big(Fu\big)(\xi) = \hat{u}(\xi) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathrm{e}^{-i\langle \xi, x \rangle} u(x)\, \mathrm{d}x,$$

where i is the imaginary unit and $\langle \cdot\,, \cdot \rangle$ is the Euclidean inner product.

Similarly, the inverse Fourier transform *is defined by*

$$\big(F^{-1}v\big)(x) = \check{v}(x) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} \mathrm{e}^{i\langle \xi, x \rangle} v(\xi)\, \mathrm{d}\xi,$$

Proposition 6.22.

- *$F : S(\mathbb{R}^d;\mathbb{C}) \to S(\mathbb{R}^d;\mathbb{C})$ and $F^{-1} : S(\mathbb{R}^d;\mathbb{C}) \to S(\mathbb{R}^d;\mathbb{C})$;*
- *F, F^{-1} are bounded linear operators on $S(\mathbb{R}^d;\mathbb{C})$;*
- *$u = \big(F^{-1} \circ F\big)u$ for all $u \in S(\mathbb{R}^d;\mathbb{C})$;*
- *If $u, v \in S(\mathbb{R}^d;\mathbb{C})$, then*

$$\langle Fu, Fv \rangle_{L^2} = \langle u, v \rangle_{L^2}. \tag{6.2}$$

See Exercises 6.5 and 6.6.

Remark 6.23. *By a limiting procedure, the details of which we omit, the Fourier transform can be extended to functions in $L^2(\mathbb{R}^d;\mathbb{C})$. The following proposition (to prove in Exercise 6.9) and lemma, which we do not prove here, will be useful in Chapter 9.*

Proposition 6.24. *The Fourier transform maps L^1 to L^∞ and L^2 to L^2. As a consequence, it maps L^q to L^p for any conjugate pair (p,q) with $q \in [1,2]$, and there exists a constant $C = C(p) > 0$ such that, for all $q \in [1,2]$,*

$$\|\hat{u}\|_{L^p} \leq C(p)\|u\|_{L^q}.$$

The same is true of the inverse Fourier transform: there exists a constant $C = C(p) > 0$ such that, for all $q \in [1,2]$,

$$\|u\|_{L^p} \leq C(p)\|\hat{u}\|_{L^q}.$$

Furthermore

$$\|u\|_{L^2} = \|\hat{u}\|_{L^2}.$$

Integrating by parts and using the fact that $u \in S(\mathbb{R}^d,\mathbb{C})$, it can be shown that

$$(F\partial^\alpha u)(\xi) = i^{|\alpha|}\prod_{j=1}^{d}\xi_j^{\alpha_j}(Fu)(\xi).$$

Thus, we have the following lemma.

Lemma 6.25. *The $H^k(\mathbb{R}^d;\mathbb{C})$ norm, $\|u\|_{H^k}$, for a given u is equivalent to*

$$\left(\int_{\mathbb{R}^d}(1+\|\xi\|_2^2)^k|(Fu)(\xi)|^2\,\mathrm{d}\xi\right)^{1/2}.$$

The proof at this level of generality is the subject of Exercise 6.8. Here we give the proof in dimension $d = 1$.

Proof (*Sketch in case $d = 1$*) Using integration by parts (first in the Schwarz space, then using approximation to go beyond this setting) shows that

$$\left(F\frac{d^\alpha u}{dx^\alpha}\right)(\xi) = (i\xi)^\alpha\big(Fu\big)(\xi).$$

Thus, using the fact that the Fourier transform preserves the L^2 norm (Proposition 6.24),

$$\left\|\frac{d^\alpha u}{dx^\alpha}\right\|_{L^2}^2 = \int_{\mathbb{R}}|\xi|^{2\alpha}|(Fu)(\xi)|^2\mathrm{d}\xi.$$

It follows that

$$\|u\|_{H^k}^2 = \sum_{0\le\alpha\le k} \left\| \frac{d^\alpha u}{dx^\alpha} \right\|_{L^2}^2 = \int_{\mathbb{R}} \Big(\sum_{0\le\alpha\le k} |\xi|^{2\alpha} \Big) |(Fu)(\xi)|^2 \mathrm{d}\xi.$$

It is a property of polynomials that there exists $c \in (0,1)$ so that, for all $\xi \in \mathbb{R}$,

$$c < \frac{\sum_{0\le\alpha\le k} |\xi|^{2\alpha}}{(1+|\xi|^2)^k} < c^{-1}.$$

Consequently

$$c < \frac{\|u\|_{H^k}^2}{\int_{\mathbb{R}} (1+\|\xi\|_2^2)^k |(Fu)(\xi)|^2 \,\mathrm{d}\xi} < c^{-1}$$

establishing that the norms are equivalent. □

Exercises

6.1 This exercise refers to the concepts of compactness, from Lecture 10, and orthogonality, from Lecture 11, and is best tackled after those chapters have been covered.

(a) Show that the family of functions $\{\varphi_j\}_{j\in\mathbb{N}}$ defined by $\varphi_j(x) = \sqrt{2}\sin(j\pi x)$ are orthonormal with respect to the inner product

$$\langle u, v\rangle_{L^2} = \int_0^1 u(x)v(x)\mathrm{d}x.$$

(b) Let $D = (0,1)$. In what follows you may assume that these functions $\{\varphi_j\}_{j\in\mathbb{N}}$ form an orthonormal basis for

$$L^2(D;\mathbb{R}) := \left\{ u : (0,1) \to \mathbb{R} \;\middle|\; \int_0^1 u(x)^2 \mathrm{d}x < \infty \right\}$$

with norm $\|u\|_{L^2}^2 = \int_0^1 u(x)^2\mathrm{d}x$. We also define the space

$$H_0^1(D;\mathbb{R}) := \left\{ u : (0,1) \to \mathbb{R} \;\middle|\; \int_0^1 u'(x)^2\mathrm{d}x < \infty,\ u(0) = u(1) = 0 \right\}$$

with norm $\|u\|_{H_0^1}^2 = \int_0^1 u'(x)^2\mathrm{d}x$. Define an inner product which induces the norm on $H_0^1(D;\mathbb{R})$ (and show that indeed your definition is a valid inner product).

(c) Show that $\{\varphi_j\}_{j\in\mathbb{N}}$ are eigenfunctions of the differential operator $-\frac{d^2}{dx^2}$ on $(0,1)$ with homogeneous Dirichlet boundary conditions. Using this fact, find expressions for the $L^2(D;\mathbb{R})$ and $H_0^1(D;\mathbb{R})$ norms in terms of the coefficients of an expansion of u in the orthonormal basis $\{\varphi_j\}_{j\in\mathbb{N}}$. Deduce that $H_0^1(D;\mathbb{R})$ is compactly embedded in $L^2(D;\mathbb{R})$.

6.2 Let D be a bounded open set in $\mathbb{R}^d$. Prove that, when equipped with the supremum norm

$$\|u\|_\infty = \sup_{x\in\overline{D}} |u(x)| = \sup_{x\in D} |u(x)|,$$

the space of continuous real-valued functions $C(\overline{D};\mathbb{R})$ is complete.

6.3 Show that the supremum norm does not define a norm on the space $C(D;\mathbb{R})$, and so in particular, $C(D;\mathbb{R})$ is not complete with respect to the supremum norm.

6.4 Show that the object defined in (4.5) is indeed a real-valued inner product.

6.5 Show that if $u \in S(\mathbb{R}^d;\mathbb{C})$, then $\hat{u}, \check{u} \in S(\mathbb{R}^d;\mathbb{C})$ (see Proposition 6.22).

6.6 Prove the identity (6.2). Explain why the identity still holds on $L^2(\mathbb{R}^d;\mathbb{C})$, and deduce that the mapping $u \mapsto \hat{u}$ is a unitary map from $L^2(\mathbb{R}^d;\mathbb{C})$ to itself (see Proposition 6.22).

6.7 Prove that the following inequality holds if and only if $k > d/2$ (here $|\cdot|$ denotes the usual Euclidean norm):

$$\int_{\mathbb{R}^d} \frac{1}{(1+|\xi|^2)^k}\,\mathrm{d}\xi < \infty\,.$$

6.8 Let $u \in S(\mathbb{R}^d;\mathbb{C})$. The Fourier transform can be viewed as a linear isomorphism $F\colon S(\mathbb{R}^d;\mathbb{C}) \to S(\mathbb{R}^d;\mathbb{C})$, i.e. a linear operator that preserves the norm (the last part is addressed in (d)). This exercise establishes some of its properties.

Note: as mentioned in the lecture, F can be extended to a bounded linear operator $F\colon L^2(\mathbb{R}^d;\mathbb{C}) \to L^2(\mathbb{R}^d;\mathbb{C})$ via the "density argument" (i.e. using the fact that $S(\mathbb{R}^d;\mathbb{C})$ is dense in $L^2(\mathbb{R}^d;\mathbb{C})$). Density is introduced in Lecture 8.

(a) Show that

$$(F[\partial^\alpha u])(\xi) = i^{|\alpha|} \prod_{j=1}^{d} \xi_j^{\alpha_j} (Fu)(\xi).$$

(b) Prove Lemma 6.25: show that the $H^k(\mathbb{R}^d;\mathbb{C})$ norm is equivalent to

$$\left(\int_{\mathbb{R}^d} (1+\|\xi\|_2^2)^k |(Fu)(\xi)|^2 \,\mathrm{d}\xi\right)^{1/2}.$$

(c) Recall the definition of the inverse Fourier transform $F^{-1}\colon S(\mathbb{R}^d;\mathbb{C}) \to S(\mathbb{R}^d;\mathbb{C})$ and note that indeed $F^{-1}\circ F = I$ on $S(\mathbb{R}^d;\mathbb{C})$. The inverse FT may again be extended to $L^2(\mathbb{R}^d;\mathbb{C})$. Using the inverse FT to express $\|u\|_{L^\infty}$ in terms of $F[u]$, show that if $u \in H^k(\mathbb{R}^d;\mathbb{C})$ and $k > d/2$, then there is a constant $C = C(k,d)$ such that

$$\|u\|_{L^\infty} \leqslant C\|u\|_{H^k}.$$

(d) Prove a version of the *Plancherel theorem*, i.e., show that the FT preserves the L^2-norm: if $u \in S(\mathbb{R}^d;\mathbb{C})$, then

$$\|Fu\|_{L^2} = \|u\|_{L^2}.$$

From this result, it can be deduced that also $\langle Fu, Fv\rangle_{L^2} = \langle u, v\rangle_{L^2}$ for any $u, v \in S(\mathbb{R}^d;\mathbb{C})$. This latter identity is also sometimes referred to as the Plancherel theorem. Alternatively, you can first prove the inner product identity and then arrive at the one above.

6.9 Prove Proposition 6.24. You may use the following theorem without proof.

Theorem 6.26 (Riesz-Thorin Interpolation Theorem)**.** *Let p_0, p_1, q_0, $q_1 \in [1,\infty]$, and let $T\colon L^{q_0} + L^{q_1} \to L^{p_0} + L^{p_1}$ be a linear operator. Assume that T is bounded from L^{q_0} to L^{p_0} and from L^{q_1} to L^{p_1}. Then T is bounded from L^{q_t} to L^{p_t} for any $t \in [0,1]$, where*

$$\frac{1}{q_t} = \frac{1-t}{q_0} + \frac{t}{q_1}, \quad \frac{1}{p_t} = \frac{1-t}{p_0} + \frac{t}{p_1}.$$

6.10 For $p \in [1,\infty)$, define the space $L^p(D;\mathbb{R}^n)$ by

$$L^p(D;\mathbb{R}^n) = \left\{ u\colon D \to \mathbb{R}^n \;\middle|\; u \text{ is measurable and } \int_D \|u(x)\|_{\mathbb{R}^n}^p \,\mathrm{d}x < \infty \right\}$$

equipped with the norm

$$\|u\|_{L^p(D;\mathbb{R}^n)} = \left(\int_D \|u(x)\|_{\mathbb{R}^n}^p \,\mathrm{d}x\right)^{\frac{1}{p}}.$$

Additionally, define the space $L^\infty(D;\mathbb{R}^n)$ by

$$L^\infty(D;\mathbb{R}^n) = \left\{ u\colon D \to \mathbb{R}^n \;\middle|\; u \text{ is measurable and } \operatorname*{ess\,sup}_{x\in D} \|u(x)\|_{\mathbb{R}^n} < \infty \right\}$$

equipped with the norm

$$\|u\|_{L^\infty(D;\mathbb{R}^n)} = \operatorname*{ess\,sup}_{x\in D} \|u(x)\|_{\mathbb{R}^n}.$$

Show that $L^p(D;\mathbb{R}^n)$ is a Banach space for $p \in [1,\infty]$, and that it is separable for $p \in [1,\infty)$. Additionally, define an inner product that makes $L^2(D;\mathbb{R}^n)$ a Hilbert space.

6.11 Let $u, v \in H^1(D;\mathbb{R})$. Show that we may write

$$\begin{aligned}\langle u, v\rangle_{H^1(D;\mathbb{R})} &= \int_D u(x)v(x)\,\mathrm{d}x + \int_D \langle \nabla u(x), \nabla v(x)\rangle_{\mathbb{R}^d}\,\mathrm{d}x \\ &= \langle u, v\rangle_{L^2(D;\mathbb{R})} + \langle \nabla u, \nabla v\rangle_{L^2(D;\mathbb{R}^d)}.\end{aligned}$$

Suppose now that $u, v \in H^2(D;\mathbb{R})$. Find a similar expression for $\langle u, v\rangle_{H^2(D;\mathbb{R})}$ using the Hessian matrices of u and v.

6.12 Recall the definitions of spaces of functions $C(D;\mathbb{R})$, $C^k(D;\mathbb{R})$, $C^\infty(D;\mathbb{R})$ and $C_c^\infty(D;\mathbb{R})$. We now introduce the following space:

$$L^p_{\mathrm{loc}}(D;\mathbb{R}) := \left\{u\colon D \to \mathbb{R} \;\middle|\; u \in L^p(K;\mathbb{R}) \text{ for any compact } K \subset D\right\}.$$

Here we also introduce the weak derivatives (cf. Definition 6.11); notice that the two definitions differ slightly because we allow D to be unbounded here (hence the use of L^1_{loc}), whereas in the lecture we only consider bounded D.

Definition 6.27. *Let $u, v \in L^1_{loc}(D;\mathbb{R})$ and α be a multi-index. Then v is called the α-th* weak partial derivative *of u if*

$$\int_D v(x)\varphi(x)\,\mathrm{d}x = (-1)^{|\alpha|}\int_D u(x)\partial^\alpha\varphi(x)\,\mathrm{d}x \quad \textit{for any } \varphi \in C_c^\infty(D;\mathbb{R}).$$

We write: $v := \partial^\alpha u$; this can be considered abuse of notation since it is not in general equal to the usual derivative, yet the notation is exactly the same.

Find (or show that it does not exist):

(a) the weak derivative of

$$u(x) = \begin{cases} 0, & x < 0, \\ x, & 0 < x < 1, \\ 1, & 1 < x; \end{cases}$$

(b) the weak second derivative of

$$u(x) = \begin{cases} x^2, & x < 1, \\ 1, & 1 < x; \end{cases}$$

(c) the weak second derivative of

$$u(x) = \|x\|_2^2, \quad x \in \mathbb{R}^d.$$

In one-dimensional examples above, let $D = \mathbb{R}$, and in part (c), let $D = \mathbb{R}^d$.

Hint: You may want to read about *Sobolev embedding theorem* which helps prove that certain weak derivatives do not exist; you only need the most minimal version. It is also possible to prove it without the theorem.

6.13 Define the Hilbert space $L^2 = L^2(I; \mathbb{C})$, where $I = (0, 1)$, via the inner product and norm

$$\langle u, v \rangle_{L^2} = \int_0^1 \overline{u(x)} v(x) dx, \quad \|u\|_{L^2}^2 = \int_0^1 |u(x)|^2 dx,$$

and consider the functions $\{\varphi^{(n)}\}$ defined by $\varphi^{(n)}(x) = \exp(2\pi i n x)$ for all $n \in \mathbb{Z}$ and $x \in I$. Show that

$$\langle \varphi^{(m)}, \varphi^{(n)} \rangle_{L^2} = \delta_{mn}.$$

Make the assumption that the $\{\varphi^{(n)}\}_{n\in\mathbb{Z}}$ form a basis for L^2 so that, for any $f \in L^2$, there is a sequence $\{\hat{f}_n\}_{n\in\mathbb{Z}}$ with the property that

$$f = \sum_{n\in\mathbb{Z}} \hat{f}_n \varphi^{(n)}$$

and the partial sums

$$f^N = \sum_{|n|\le N} \hat{f}_n \varphi^{(n)}$$

converge to f in L^2. Define $\ell^2 = \ell^2(\mathbb{Z}; \mathbb{C})$, the space of functions from $\mathbb{Z}$ into $\mathbb{C}$ (complex-valued sequences over the positive and negative integers), via the inner product and norm

$$\langle u, v \rangle_{\ell^2} = \sum_{n\in\mathbb{Z}} \overline{u_n} v_n, \quad \|u\|_{\ell^2}^2 = \sum_{n\in\mathbb{Z}} |u_n|^2.$$

Then define operator S by

$$Sf = \{\hat{f}_n\}_{n\in\mathbb{Z}}.$$

Show that S is linear, that $S : L^2 \to \ell^2$ and that

$$\|Sf\|_{\ell^2} = \|f\|_{L^2}.$$

7
Linear Operators

Linear operators between normed vector spaces are a key concept throughout these lectures, underpinning important ideas such as duality, continuous embedding and compact operators. In this lecture we introduce the basic theory underpinning linear operators. Note that linear *operators* are sometimes termed linear *maps* or linear *functions*.

7.1 Bounded Linear Operators

By $\mathbb{K}^{m\times n}$ we will denote all $m \times n$ matrices with entries in $\mathbb{K}$, where $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. Every element of $\mathbb{K}^{m\times n}$ can be uniquely identified with a vector in $\mathbb{K}^{mn}$ (and vice versa). Then, we can define a norm on $\mathbb{K}^{mn}$ as a norm on $\mathbb{K}^{m\times n}$.

Examples 7.1. Let $A \in \mathbb{R}^{m\times n}$.

- $\|A\|_{\max} = \max\limits_{1\le i\le m, 1\le j\le n} |a_{ij}|$ is called the *maximum matrix* norm.
- $\|A\|_F = \left(\sum_{i,j=1}^{m,n} |a_{ij}|^2\right)^{1/2}$ is called the *Frobenius* or *Hilbert–Schmidt* norm.

◊

However, matrices have additional structure that vectors do not have: if $A_1, A_2 \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$, then $A_1A_2 \in \mathbb{R}^{n\times n}$ and $A_iv \in \mathbb{R}^n$, $i = 1, 2$, are defined. It is natural, and useful, to require

$$\|A_1A_2\| \le \|A_1\|\,\|A_2\|, \quad \|Av\| \le \|A\|\,\|v\|.$$

The norms in the example above do not satisfy these inequalities. However, there is a useful way to construct matrix norms on $\mathbb{R}^{n\times n}$, that

do satisfy these inequalities: by *inducing* them from the underlying norm on the vector space $\mathbb{R}^n$ that the matrix maps between. In fact the idea of *induced norms* applies more generally for linear transformations between any normed vector space and so we now adopt this setting.

Let V, W be normed vector spaces when equipped with norms $\|\cdot\|_V, \|\cdot\|_W$, respectively.

Definition 7.2. *An operator $L\colon V \to W$ is* linear *if*

$$L(\alpha v_1 + \beta v_2) = \alpha L v_1 + \beta L v_2 \quad \text{for all } \alpha, \beta \in \mathbb{K},\ v_1, v_2 \in V.$$

Definition 7.3. *An operator $L\colon V \to W$ is* continuous *if for any open $A \subseteq W$, it follows that $L^{-1}(A) \subseteq V$ is open.*

Definition 7.4. *An operator $L\colon V \to W$ is* continuous at a point x_0 *if for every $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) > 0$ such that for any $y \in V$ satisfying $\|x_0 - y\|_V < \delta$, it holds that*

$$\|Lx_0 - Ly\|_W < \varepsilon.$$

Remark 7.5. *It is true that if L is continuous at every point in V, then it is continuous. Moreover, an even stronger statement is true for linear operators: if L is linear and continuous at 0 (or any other point), then it is continuous (see Exercise 7.3).*

Definition 7.6. *An operator $L\colon V \to W$ is* bounded *if there exists $K \in \mathbb{R}^+$ such that*

$$\|Lv\|_W \leq K\|v\|_V \quad \forall v \in V. \tag{7.1}$$

Let $\mathcal{L}(V, W)$ denote the space of all bounded linear operators from V to W. Define

$$\|L\|_{\mathcal{L}(V,W)} := \sup_{\|v\|_V = 1} \|Lv\|_W = \sup_{v \neq 0} \frac{\|Lv\|_W}{\|v\|_V}. \tag{7.2}$$

The following proposition establishes that with the above norm, we thereby define a normed vector space of linear operators and is proved as Exercise 7.1. The norm (7.2) on linear operators from V into W is said to be the norm *induced* by the norms on V and W; the *induced norm* or *operator norm*, for short.

Proposition 7.7. *The space $\mathcal{L}(V, W)$ equipped with norm $\|\cdot\|_{\mathcal{L}(V,W)}$ is a normed vector space.*

Example 7.8. The fact that both spaces V and W are explicitly stated in the definition of the operator norm (7.2) is not just a notational detail; it is an essential part of the definition of the operator norm. To see this, consider the right-shift operator S_r acting on infinite sequences, defined by

$$S_r u = (0, u_1, u_2, u_3, \ldots),$$

for any $u = (u_1, u_2, u_3, \ldots)$. Clearly, S_r is bounded when viewed as an operator from $\ell^p(\mathbb{N};\mathbb{R})$ to $\ell^p(\mathbb{N};\mathbb{R})$; however, it is not bounded when viewed as mapping from $\ell^2(\mathbb{N};\mathbb{R})$ to $\ell^1(\mathbb{N};\mathbb{R})$, for example; this is because it is not possible to bound the $\ell^1(\mathbb{N};\mathbb{R})$ norm of a sequence by the $\ell^2(\mathbb{N};\mathbb{R})$ norm – to see this simply construct a sequence which is in $\ell^2(\mathbb{N};\mathbb{R})$ but not in $\ell^1(\mathbb{N};\mathbb{R})$. The operator is bounded when viewed as mapping from $\ell^1(\mathbb{N};\mathbb{R})$ to $\ell^2(\mathbb{N};\mathbb{R})$, however; indeed Example 9.6 in Lecture 9 shows that it has norm less than 1. ◊

Lemma 7.9.

$$\|L\|_{\mathcal{L}(V,W)} = \inf\{K \in \mathbb{R}^+ \mid (7.1) \text{ holds}\}. \tag{7.3}$$

Furthermore

$$\|Lu\|_W \leq \|L\|_{\mathcal{L}(V,W)} \|u\|_V.$$

Proof We drop the suffices V, W on the vector space norms as the context determines which is in play. We start from the definition of a norm. For any $u \in V$,

$$\frac{\|Lu\|}{\|u\|} \leq \sup_{v \neq 0} \frac{\|Lv\|}{\|v\|} = \|L\|_{\mathcal{L}(V,W)}. \tag{7.4}$$

Therefore, $\|Lu\| \leq \|L\|_{\mathcal{L}(V,W)} \|u\|$ for any $u \in V$. The norm $\|L\|_{\mathcal{L}(V,W)}$ is the least K that can be used in (7.1). Otherwise, there exists K such that

$$\sup_{v \neq 0} \frac{\|Lv\|}{\|v\|} \leq K < \|L\|_{\mathcal{L}(V,W)} = \sup_{v \neq 0} \frac{\|Lv\|}{\|v\|},$$

which is a contradiction. The inequality follows from equation 7.4. □

Example 7.10. Let $A \in \mathbb{R}^{n \times m}$. Then the following defines an induced norm:

$$\|A\|_{m \to n} := \sup_{v \neq 0} \frac{\|Av\|_{\mathbb{R}^n}}{\|v\|_{\mathbb{R}^m}},$$

where $\|\cdot\|_{\mathbb{R}^n}$, $\|\cdot\|_{\mathbb{R}^m}$ are any norms on $\mathbb{R}^n$, $\mathbb{R}^m$, respectively. In particular,

when $n = m$ and $\|\cdot\|_{\mathbb{R}^n}$ is chosen to be a p-norm $\|\cdot\|_p$, $p \in [1, \infty]$, we then write

$$\|A\|_p := \sup_{v \neq 0} \frac{\|Av\|_p}{\|v\|_p}.$$

◊

Proposition 7.11. *Equation* (7.2) *defines a norm on* $\mathcal{L}(V, W)$ *and furthermore, if* $L_1 : U \to V$, $L_2 : V \to W$, *then* $L_2 L_1 : U \to W$ *satisfies*

$$\|L_2 L_1\|_{\mathcal{L}(U,W)} \leq \|L_2\|_{\mathcal{L}(V,W)} \|L_1\|_{\mathcal{L}(U,V)}.$$

Proof We drop the suffices U, V, W on the vector space norms as the context determines which is in play. The fact that $\|\cdot\|_{\mathcal{L}(V,W)}$ is a norm on $\mathcal{L}(V, W)$ is a simple consequence of respective norm axioms of the $\|\cdot\|_V$ and $\|\cdot\|_W$ norms. For the last part,

$$\begin{aligned} \|L_2 L_1\|_{\mathcal{L}(U,W)} &= \sup_{u \neq 0} \frac{\|L_2 L_1 u\|}{\|u\|} \\ &\leq \sup_{u \neq 0} \frac{\|L_2\|_{\mathcal{L}(V,W)} \|L_1 u\|}{\|u\|} \\ &= \|L_2\|_{\mathcal{L}(V,W)} \|L_1\|_{\mathcal{L}(U,V)}. \end{aligned}$$

□

Example 7.12. Let $V = W = \mathbb{R}^n$. Then

$$\|A\|_\infty = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

To see this note that

$$\|Ax\|_\infty = \max_{1 \leq i \leq n} |(Ax)_i| = \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} a_{ij} x_j \right| \leq \max_{1 \leq i \leq n} \left(\sum_{j=1}^{n} |a_{ij}| \right) \|x\|_\infty.$$

Therefore,

$$\|A\|_\infty = \sup_{x \neq 0} \frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.$$

It only remains to find x for which the last inequality turns into equality. We will assume $A \neq 0$ since this case is trivial. Choose $k \in \{1, 2, \ldots, n\}$ such that

$$\max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}| = \sum_{j=1}^{n} |a_{kj}|,$$

and define $x \in \mathbb{R}^n$ by $x_j := a_{kj}/|a_{kj}|$ for all $j = 1, \ldots, n$, with the convention that this is 0 when a_{kj} is 0. Then $\|x\|_\infty = 1$ and

$$\begin{aligned}\|A\|_\infty &= \max_{\|y\|_\infty = 1} \|Ay\|_\infty \\ &\geq \|Ax\|_\infty = \max_{1 \leq i \leq n} \left| \sum_{j=1}^{n} a_{ij} \frac{a_{kj}}{|a_{kj}|} \right| \\ &\geq \left| \sum_{j=1}^{n} a_{kj} \frac{a_{kj}}{|a_{kj}|} \right| = \sum_{j=1}^{n} |a_{kj}| = \max_{1 \leq i \leq n} \sum_{j=1}^{n} |a_{ij}|.\end{aligned}$$

This is the required result. ◊

Definition 7.13. *Let $L \in \mathcal{L}(V, W)$. Assume that there exist m linearly independent elements in W, $\{\varphi^{(j)}\}_{j=1}^m$, with the property that, for every $v \in V$, $Lv \in \operatorname{span}\{\varphi^{(j)}\}_{j=1}^m$. Then we say that L has* rank m.

7.2 Bounded Linear Operators Form a Banach Space

Theorem 7.14. *If V is a normed vector space, and if W is a Banach space, then $\mathcal{L}(V, W)$ with norm* (7.2) *is a Banach space.*

Proof Let $\{L_n\}$ denote a Cauchy sequence in $\mathcal{L}(V, W)$. We identify an element L of $\mathcal{L}(V, W)$ and show that $L_n \to L$ in the norm (7.2). By the Cauchy property of $\{L_n\}$, we know that, for any $\varepsilon > 0$, there is an integer $N = N(\varepsilon)$ such that for all $n, m \geq N$,

$$\|L_n - L_m\|_{\mathcal{L}(V,W)} < \varepsilon.$$

By the definition of the norm on linear maps we deduce that, for any $v \in V$ and for all $n, m \geq N$,

$$\|(L_n v - L_m v)\|_W = \|(L_n - L_m)v\|_W < \varepsilon \|v\|_V, \tag{7.5}$$

for the same choice of $N = N(\varepsilon)$. This proves that the sequence $\{L_n v\}$ is Cauchy, and hence convergent, in W. Thus, for each $v \in V$ we may define operator L acting on element $v \in V$ by

$$Lv := \lim_{n \to \infty} L_n v.$$

To complete the proof, we need to show that $L \in \mathcal{L}(V, W)$ and that L_n converges to L in the norm on $\mathcal{L}(V, W)$.

To prove that L is linear we note that

$$\begin{aligned} L(\alpha v_1 + \beta v_2) &= \lim_{n\to\infty} L_n(\alpha v_1 + \beta v_2) \\ &= \lim_{n\to\infty} \big(\alpha L_n v_1 + \beta L_n v_2\big) \\ &= \alpha \lim_{n\to\infty} L_n v_1 + \beta \lim_{n\to\infty} L_n v_2 \\ &= \alpha L v_1 + \beta L v_2. \end{aligned}$$

To prove that L is bounded take the limit $m \to \infty$ in (7.5) to obtain

$$\|L_n v - Lv\|_W < \varepsilon \|v\|_V. \tag{7.6}$$

Fixing an n in this expression shows that $(L - L_n)$ is a bounded linear map. Thus, since $L_n \in \mathcal{L}(V, W)$, $L = (L - L_n) + L_n$ is also a bounded linear map. Thus L is an element of $\mathcal{L}(V, W)$. In addition, taking the limit $n \to \infty$ in (7.6), noting that $\varepsilon > 0$ is arbitrary, shows that $\|L_n - L\|_{\mathcal{L}(V,W)} \to 0$. □

7.3 Banach Algebra

Definition 7.15. *A* Banach algebra $(\mathcal{X}, \|\cdot\|)$ *over* $\mathbb{K}$ *is a Banach space over* $\mathbb{K}$ *with the additional multiplicative structure that*

$$AB \in \mathcal{X} \qquad \textit{for all} \qquad A, B \in \mathcal{X}.$$

Furthermore, for all $A, B, C \in \mathcal{X}$ *and* $\alpha \in \mathbb{K}$*, the following hold:*

- $(AB)C = A(BC), A(B + C) = AB + AC, (B + C)A = BA + CA$;
- $\alpha(AB) = (\alpha A)B = A(\alpha B), \|AB\| \le \|A\|\|B\|$.

Finally, there exists an element $E \in \mathcal{X}$ *such that* $A = AE = EA$ *and* $\|E\| = 1$.

Example 7.16. If $(X, \|\cdot\|)$ is a Banach space over $\mathbb{K}$, then recall that $\mathcal{X} = \mathcal{L}(X, X)$ is a Banach space when equipped with the norm

$$\|T\| = \sup_{\|u\|=1} \|Tu\|. \tag{7.7}$$

It is immediate from this definition that, for all $v \in X$,

$$\|Tv\| \le \|T\|\|v\|. \tag{7.8}$$

In fact, $(\mathcal{X}, \|\cdot\|)$ is a Banach algebra, with AB defined by

$$(AB)u \coloneqq A(Bu) \quad \text{for all} \quad u \in X$$

and E being the identity operator $Eu = u$ for all $u \in X$. The only thing to check is that $\|AB\| \leq \|A\|\|B\|$. To see this, note that, by (7.7) and (7.8),

$$\|AB\| = \sup_{\|u\|=1} \|A(Bu)\| \leq \sup_{\|u\|=1} \|A\|\|Bu\| = \|A\|\|B\|.$$

◊

Lemma 7.17. *Let $\mathcal{X}$ be a Banach algebra, and let $A, B \in \mathcal{X}$ and sequences $\{A_n\}, \{B_n\}$ in $\mathcal{X}$ for all $n \in \mathbb{N}$. Then*

- $\|A^k\| \leq \|A\|^k$ *for all* $k \in \mathbb{N}$;
- *If $A_n \to A$ and $B_n \to B$ in $\mathcal{X}$, then $A_n B_n \to AB$ in $\mathcal{X}$ as $n \to \infty$.*

Proof For the first item note that $\|A^{m+1}\| = \|A^m A\| \leq \|A^m\|\|A\|$ and use induction. For the second, note that convergent sequences are bounded and hence that there is positive $K < \infty$ such that

$$\sup_n \|A_n\| \leq K, \quad \sup_n \|B_n\| \leq K,$$

and furthermore this implies that $\|A\| \leq K$ and $\|B\| \leq K$. Thus, we get

$$\begin{aligned} \|A_n B_n - AB\| &= \|(A_n - A)B_n - A(B - B_n)\| \\ &\leq \|A_n - A\|\|B_n\| + \|A\|\|B - B_n\| \\ &\leq K(\|A_n - A\| + \|B - B_n\|). \end{aligned}$$

The result follows from the convergence of $\{A_n\}$ and $\{B_n\}$. □

7.4 Outer Products

Recall Definition 4.37 here specified to the Hilbert space setting:

Definition 7.18. *Given $a, b \in H$, we define the* outer product $a \otimes b \in \mathcal{L}(H, H)$ *by*

$$(a \otimes b)c = \langle b, c \rangle a$$

for any $c \in H$.

The statement that $a \otimes b \in \mathcal{L}(H, H)$ is justified by the following:

Proposition 7.19.

$$\|a \otimes b\|_{\mathcal{L}(H,H)} = \|a\|\|b\|.$$

Furthermore $a \otimes b$ is a rank one operator.

Proof By definition of the operator norm and by the Cauchy-Schwarz inequality:

$$\begin{aligned}\|a \otimes b\|_{\mathcal{L}(H,H)} &= \sup_{x\neq 0} \frac{\|\langle b, x\rangle a\|}{\|x\|} \\ &= \sup_{x\neq 0} \frac{|\langle b, x\rangle|\,\|a\|}{\|x\|} \\ &= \|a\| \sup_{x\neq 0} \frac{|\langle b, x\rangle|}{\|x\|} \\ &\leq \|a\|\|b\|\end{aligned}$$

Furthermore, choosing $x = b$ shows that the upper-bound is attained. The rank property follows form the fact that $(a \otimes b)c \in \operatorname{span}\{a\}$ for all $c \in H$. □

Exercises

7.1 Show that the space of bounded linear maps from V to W is a vector space and that (7.3) does indeed define a norm on this space.

7.2 Prove that (7.3) is equivalent to the original definition (7.2) of the operator norm.

7.3 Let $A\colon X \to Y$ be a linear operator between two Banach spaces. Show that the following are equivalent:

(a) A is bounded;
(b) A is Lipschitz;
(c) A is continuous;
(d) A is continuous at 0.

7.4 Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ be normed vectors spaces, and let $T\colon X \to Y$ be a bounded linear operator. Show that for any $u \in X$ and $r > 0$,

$$\sup_{v\in B_r(u)} \|Tv\|_Y \geq r\|T\|_{\mathcal{L}(X,Y)}$$

where $B_r(u) = \{v \in X : \|u - v\|_X < r\}$ is the ball of radius r centered at u.

7.5 Let $(X, \|\cdot\|_X)$ be a Banach space and $(Y, \|\cdot\|_Y)$ a normed vector space. Let $\mathcal{F} \subseteq \mathcal{L}(X, Y)$ be a family of bounded linear operators from X to Y. Assume that for each $u \in X$,

$$\sup_{T\in\mathcal{F}} \|Tu\|_Y < \infty.$$

Show that

$$\sup_{T\in\mathcal{F}} \|T\|_{\mathcal{L}(X,Y)} < \infty.$$

Hint. Suppose that the result is not true; so, you may choose a sequence $\{T_n\}_{n=1}^{\infty} \subseteq \mathcal{F}$ with $\|T_n\|_{\mathcal{L}(X,Y)} \geq 4^n$ for each n. Use Exercise 7.4 to construct a sequence $\{u_n\} \subseteq X$ with

$$\|u_n - u_{n-1}\|_X \leq 3^{-n} \quad \text{and} \quad \|T_n u_n\|_Y \geq \frac{2}{3}\, 3^{-n} \|T_n\|_{\mathcal{L}(X,Y)}$$

for each n and derive a contradiction.

7.6 Let $(X, \|\cdot\|_X)$ be a Banach space and $(Y, \|\cdot\|_Y)$ a normed vector space. Let $\{T_n\} \subseteq \mathcal{L}(X, Y)$ be such that $\{T_n u\}$ converges in Y for every $u \in X$. Using Exercise 7.5, show that the map $T\colon X \to Y$ given by

$$Tu := \lim_{n\to\infty} T_n u$$

defines a bounded linear operator. Does it follow that $\|T_n - T\|_{\mathcal{L}(X,Y)} \to 0$?

7.7 Let $\mathbb{R}[x]$ denote the space of polynomials with real coefficients, equipped with the norm

$$\left\| \sum_{j=0}^{m} a_j x^j \right\| = \max_{0\leq j\leq m} |a_j|.$$

(a) Verify that this defines a norm on $\mathbb{R}[x]$.

(b) Define the family of maps $T_n\colon \mathbb{R}[x] \to \mathbb{R}$ by

$$T_n\left(\sum_{j=0}^{m} a_j x^j \right) = \sum_{j=0}^{n} a_j.$$

Using Exercise 7.5, show that $\mathbb{R}[x]$ is not complete with respect to this norm.

7.8 Let $A, B \in \mathcal{L}(X, X)$ be bounded operators on a Banach space X such that $AB = BA$. Show that for any $n \in \mathbb{N}$,

$$(A + B)^n = \sum_{k=0}^{n} \binom{n}{k} A^k B^{n-k}.$$

7.9 Let X, Y be Banach spaces. We will say that a map $f\colon X \to Y$ is open if $f(U) \subseteq Y$ is open for any open $U \subset X$. A key result related to open maps is the *open mapping theorem*:

Theorem 7.20. *If X, Y are Banach spaces and $T \in \mathcal{L}(X, Y)$ is surjective, then T is open.*

You may use this result without proof in what follows.

(a) Prove that if $T \in \mathcal{L}(X, Y)$ is invertible, then $T^{-1} \in \mathcal{L}(Y, X)$.

(b) Given a map $T\colon X \to Y$, define its graph $G(T) \subseteq X \times Y$ by

$$G(T) = \{(u, v) \in X \times Y : v = Tu\}.$$

Equip $X \times Y$ with the product norm $\|(u, v)\|_{X\times Y} = \|u\|_X + \|v\|_Y$. Prove that if $T\colon X \to Y$ is a linear operator such that $G(T)$ is closed, then $T \in \mathcal{L}(X, Y)$.

Hint: for part (a), use Exercise 7.3.

7.10 Let $\ell^0 \subset \mathbb{R}^\infty$ be the space of sequences with finitely many non-zero entries, equipped with the norm $\|\cdot\|_{\ell^2}$. Define the map $T\colon \ell^0 \to \ell^0$ by $(Tu)_j = u_j/j$ for each j. Show that T is linear, bounded and invertible, and that T^{-1} is unbounded. Does this contradict Exercise 7.9(a)?

7.11 Let X be a Banach space when equipped with either of the norms $\|\cdot\|_1$ or $\|\cdot\|_2$. Suppose that there exists $C > 0$ such that

$$\|u\|_1 \leq C\|u\|_2 \quad \text{for all } u \in X.$$

Using Exercise 7.9(a), show that $\|\cdot\|_1$ and $\|\cdot\|_2$ are equivalent.

7.12 Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and let $T, S\colon H \to H$ be linear operators. Suppose that

$$\langle Tu, v\rangle = \langle u, Sv\rangle \quad \text{for all } u, v \in H.$$

Using Exercise 7.9(b), show that both $T, S \in \mathcal{L}(H, H)$. Deduce that, in particular, any everywhere-defined symmetric operator is bounded (for symmetric operator, see Definition 11.13).

7.13 Define the functional $\delta\colon C([0,1]) \to \mathbb{R}$ by $\delta(u) = u(0)$. Calculate the operator norm of δ when $C([0,1])$ is equipped with the L^p norm for any $p \in [1, \infty]$.

7.14(a) Define $\delta\colon C([0,1]) \to \mathbb{R}$ by $\delta(u) = u(0)$, where $C([0,1])$ is equipped with the L^∞ norm. Show that there exists a bounded linear functional $\overline{\delta}\colon L^\infty([0,1]) \to \mathbb{R}$ that extends δ.
Hint: Use Theorem 8.11.

(b) Define $u_n\colon [0,1] \to \mathbb{R}$ by $u_n(x) = e^{-nx}$ for each $n \in \mathbb{N}$. Show that there does not exist $g \in L^1([0,1])$ such that for all $n \in \mathbb{N}$,

$$\overline{\delta}(u_n) = \int_0^1 u_n(x)g(x)\,dx.$$

(c) Deduce that there is no $g \in L^1([0,1])$ that represents $\overline{\delta}$, and so formally, $(L^\infty)^* \supsetneq L^1$.

7.15 Let $(H, \langle \cdot, \cdot \rangle)$ be a Hilbert space, and let $u, v \in H$. Define the tensor product $u \otimes v : H \to H$ by

$$(u \otimes v)(w) = \langle v, w \rangle u \quad \text{for all } w \in H.$$

Show that $u \otimes v \in \mathcal{L}(H, H)$ with $\|u \otimes v\|_{\mathcal{L}(H,H)} = \|u\|\|v\|$.

7.16 Find examples of 2×2 matrices which demonstrate that the Frobenius and maximum norms on matrices are not induced norms. Use the fact that induced norms inherit the properties described in the section on Banach algebras in order to construct counterexamples.

8
Duality, Density and Basis

This lecture introduces the key concept of dual Banach space, defined from a given Banach space by considering the Banach space of bounded linear functionals on the given Banach spaces. We then proceed to discuss density, separability and various notions of basis in a Banach space. A key role is played by *linear functionals*: scalar-valued functions taking inputs from a normed vector space.

8.1 Duality

Example 8.1. Consider the set $\mathcal{L}$ of linear functions on $\mathbb{R}^n$. The action of any $f \in \mathcal{L}$ on element $v \in \mathbb{R}^n$ may be written, for some $w \in \mathbb{R}^n$, as

$$f(v) = \langle v, w \rangle, \tag{8.1}$$

using the Euclidean inner product. The converse is also true: any $w \in \mathbb{R}^n$ defines a linear function on $\mathbb{R}^n$ via the equation (8.1). Thus the class of linear functionals on $\mathbb{R}^n$ is equivalent to the set $\mathbb{R}^n$ itself. If we compute the norm of f as a linear map from $\mathbb{R}^n$ into $\mathbb{R}$, then it is given by

$$\|f\|_{\mathcal{L}} = \sup_{v \neq 0} \frac{|\langle v, w \rangle|}{\|v\|_2}.$$

By the Cauchy-Schwarz inequality it follows that

$$\|f\|_{\mathcal{L}} \leq \|w\|_2,$$

where $\|\cdot\|_2$ denotes the Euclidean norm. By taking $v = w$ we see that $\|f\|_{\mathcal{L}} = \|w\|_2$. We say that $\mathcal{L}$ is *isometrically isomorphic* to $\mathbb{R}^n$ because of the identification of every element of $\mathcal{L}$ with an element of $\mathbb{R}^n$ and vice

versa, together with the norm preservation identity. We may write f_w to denote the dependence of the linear functional on element $w \in \mathbb{R}^n$. ◊

Definition 8.2. *The* dual space *of a normed vector space X over $\mathbb{K}$ is the Banach space $X^* := \mathcal{L}(X, \mathbb{K})$. For any $f \in X^*$ we write*

$$\|f\|_* := \|f\|_{X^*} = \|f\|_{\mathcal{L}(X,\mathbb{K})}.$$

In particular, given any Banach space X, we may associate with it another Banach space X^*, the *dual Banach space*. To get intuition about dual spaces we study duality in ℓ^p, with our main result being Theorem 8.6 below; it is analogous to what is demonstrated in Example 8.1 in the case of $\mathbb{R}^n$ equipped with the Euclidean norm. We preface the theorem with a related lemma.

Lemma 8.3. *Let p, q be conjugate. Then any $v \in \ell^p$ defines an element of the dual space of ℓ^q. This element $f_v \in (\ell^q)^*$ is given by*

$$f_v(w) = \sum_{j=1}^{\infty} v_j w_j \tag{8.2}$$

and

$$\|f_v\|_{(\ell^q)^*} \leq \|v\|_{\ell^p}. \tag{8.3}$$

Proof It is clear that, for fixed $v \in \ell^p$, $f_v : \ell^q \to \mathbb{R}$ is well-defined and linear, by the Hölder inequality. Furthermore, this inequality shows that, for all $w \in \ell^q$,

$$|f_v(w)| \leq \|v\|_{\ell^p} \|w\|_{\ell^q}.$$

Thus

$$\|f_v\|_{(\ell^q)^*} = \sup_{w \neq 0} \frac{|f_v(w)|}{\|w\|_{\ell^q}} \leq \|v\|_{\ell^p}$$

as required. □

Definition 8.4. *Two normed vector spaces $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ are* isometrically isomorphic *if there exists an invertible linear mapping between the spaces which preserves the norm. We write $X \simeq Y$.*

Example 8.5. Consider $\mathbb{R}^n$ equipped with the Euclidean norm. Then $\left(\mathcal{L}(\mathbb{R}^n, \mathbb{R})\right)^* \simeq \mathbb{R}^n$. See Example 8.1. ◊

Theorem 8.6. *If $q \in (1, \infty)$ and p, q are conjugate indices, then $(\ell^q)^* \simeq \ell^p$.*

Proof We first show that any element of ℓ^p gives rise to an element of $(\ell^q)^*$ with the same norm. We have already demonstrated that (8.3) holds. We now demonstrate equality in this identity. To do so, we work with a particular sequence w defined by $w_k = |v_k|^p / v_k$, unless $v_k = 0$ in which case we define $w_k = 0$ too. With this definition, we have

$$|w_k|^q = |v_k|^{q(p-1)} = |v_k|^p$$

and, since $v \in \ell^p$, it follows that

$$\|w\|_{\ell^q}^q = \sum_{j=1}^{\infty} |w_k|^q = \sum_{j=1}^{\infty} |v_k|^p = \|v\|_{\ell^p}^p < \infty$$

and hence that $w \in \ell^q$. Furthermore,

$$f_v(w) = \sum_{j=1}^{\infty} v_j w_j = \sum_{j=1}^{\infty} |v_j|^p = \|v\|_{\ell^p}^p.$$

Thus

$$|f_v(w)| = \|v\|_{\ell^p}^p.$$

Also

$$|f_v(w)| \leq \|f_v\|_{(\ell^q)^*} \|w\|_{\ell^q} = \|f_v\|_{(\ell^q)^*} \|v\|_{\ell^p}^{p/q}.$$

Combining these last two displays shows that

$$\|v\|_{\ell^p}^p \leq \|f_v\|_{(\ell^q)^*} \|v\|_{\ell^p}^{p/q}$$

and since $p - p/q = 1$ we deduce that $\|v\|_{\ell^p} \leq \|f_v\|_{(\ell^q)^*}$. Combining with (8.3) demonstrates that $\|v\|_{\ell^p} = \|f_v\|_{(\ell^q)^*}$.

It remains to show that *any* element $f \in (\ell^q)^*$ can be identified with a linear functional f_v of the form (8.2) for some $v \in \ell^p$. To this end we define $e^{(k)} \in \ell^p$ by $e_j^{(k)} = \delta_{jk}$. Then any $w \in \ell^q$ can be written as

$$w = \sum_{k=1}^{\infty} w_k e^{(k)}$$

and

$$f(w) = \sum_{k=1}^{\infty} w_k f(e^{(k)}).$$

We thus define the sequence v by $v_k = f(e^{(k)})$. It may be shown that $\|v\|_{\ell^p} = \|f\|_{(\ell^q)^*}$ and hence also that $v \in \ell^p$, concluding the proof. □

Example 8.7.

If $q \in (1, \infty)$ and p, q are conjugate, then $\left(L^q(D; \mathbb{R}^n)\right)^* \simeq L^p(D; \mathbb{R}^n)$. ◊

Example 8.8. Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$. Then the dual norm is

$$\|y\|_* \coloneqq \max_{x \neq 0} \frac{|\langle y, x\rangle|}{\|x\|} = \max_{\|x\|=1} |\langle y, x\rangle|.$$

◊

Proposition 8.9. *$(\mathbb{R}^n, \|\cdot\|_1)$ and $(\mathbb{R}^n, \|\cdot\|_\infty)$ are duals of one another.*

Proof We need to show $\|\cdot\|_1$ is dual of $\|\cdot\|_\infty$, i.e.,

$$\max_{\|y\|_\infty=1} |\langle x, y\rangle| = \|x\|_1.$$

Notice

$$|\langle x, y\rangle| \leq \sum_{j=1}^n |x_j||y_j| \leq \max_{1\leq j\leq n} |y_j| \sum_{j=1}^n |x_j| = \|y\|_\infty \|x\|_1$$

hence

$$\max_{\|y\|_\infty=1} |\langle x, y\rangle| \leq \|x\|_1.$$

Now we show the bound is attained. To this end, let

$$y_j = \begin{cases} \frac{x_j}{|x_j|}, & x_j \neq 0 \\ 0, & x_j = 0. \end{cases}$$

Then clearly $\|y\|_\infty = 1$. Also,

$$|\langle x, y\rangle| = \left| \sum_{j=1}^n \frac{x_j^2}{|x_j|} \right| = \sum_{j=1}^n |x_j| = \|x\|_1$$

hence

$$\max_{\|y\|_\infty=1} |\langle x, y\rangle| = \|x\|_1$$

as desired. Now we show $\|\cdot\|_\infty$ is the dual of $\|\cdot\|_1$, i.e.,

$$\max_{\|y\|_1=1} |\langle x, y\rangle| = \|x\|_\infty.$$

Swapping x and y, we have

$$|\langle x, y\rangle| \leq \|x\|_\infty \|y\|_1$$

as above. Now we show that the bound is attained. Since $x \neq 0$, there exists $K \in \{1, \ldots, n\}$ such that $|x_K| = \|x\|_\infty > 0$. Define

$$y_j = \delta_{jK} \frac{x_K}{|x_K|}$$

so $\|y\|_1 = 1$. Then

$$\langle x, y \rangle = |x_K| = \|x\|_\infty,$$

showing that

$$\max_{\|y\|_1 = 1} |\langle x, y \rangle| = \|x\|_\infty.$$

□

Example 8.10. For $A \in \mathbb{C}^{n \times n}$, we have

$$\|A\|_1 = \|A^*\|_\infty = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |a_{ij}|.$$

To see this, note

$$\begin{aligned}
\|A\|_1 &= \max_{\|x\|_1 = 1} \|Ax\|_1 \\
&= \max_{\|x\|_1 = 1} \left(\max_{\|y\|_\infty = 1} |\langle Ax, y \rangle| \right) \\
&= \max_{\|y\|_\infty = 1} \left(\max_{\|x\|_1 = 1} |\langle Ax, y \rangle| \right) \\
&= \max_{\|y\|_\infty = 1} \left(\max_{\|x\|_1 = 1} |\langle x, A^* y \rangle| \right) \\
&= \max_{\|y\|_\infty = 1} \|A^* y\|_\infty \\
&= \|A^*\|_\infty.
\end{aligned}$$

◊

8.2 Duality, Hahn–Banach and Weak Convergence

Let X be a Banach space, with dual space X^* comprising all bounded linear functionals from X into $\mathbb{R}$ (i.e., we restrict to vector spaces over the reals).

Theorem 8.11 (Hahn–Banach Theorem)**.** *Let U be a subspace of Banach space X. Suppose that $f : U \to \mathbb{R}$ is a linear functional on U such that*

$$|f(x)| \leq M\|x\| \qquad \text{for all} \qquad x \in U$$

for some $M > 0$. Then there exists an $F \in X^*$ that extends f (i.e., $F(x) = f(x)$ for all $x \in U$) and does not increase its norm,

$$|F(x)| \leq M\|x\| \qquad \text{for all} \qquad x \in X.$$

We do not prove this here, but it has some interesting consequences including Exercise 11.7 and the following Lemma 8.12. The direct Corollary 8.13 is very useful as it shows that elements of a Banach space are completely characterized by the action of all linear functionals on them.

Lemma 8.12. *Let $x \in X$. Then there exists an $f \in X^*$ such that $\|f\|_{X^*} = 1$ and $f(x) = \|x\|$.*

Proof Define $\hat{f}$ on the linear space U spanned by x as

$$\hat{f}(\alpha x) = \alpha\|x\|.$$

Then $\hat{f}(x) = \|x\|$ and $|\hat{f}(z)| \leq \|z\|$ for all $z \in U$. Extend $\hat{f}$ to an $f \in X^*$; then $\|f\|_{X^*} = 1$ and $f(x) = \hat{f}(x) = \|x\|$. □

Corollary 8.13. *Let $x, y \in X$. If $f(x) = f(y)$ for every $f \in X^*$, then $x = y$.*

Proof Suppose $f(x) = f(y)$ and assume for contradiction that $x \neq y$. By the preceding lemma, there exists $f \in X^*$ with $\|f\|_{X^*} = 1$ such that $f(x-y) = \|x-y\| \neq 0$. But since f is linear it follows that $0 = f(x)-f(y) = \|x - y\| \neq 0$, the required contradiction. □

Definition 8.14. *A sequence $\{x^{(n)}\} \in X$ converges weakly to $x \in X$, written $x^{(n)} \rightharpoonup x$, if*

$$f(x^{(n)}) \to f(x) \qquad \text{for all} \qquad f \in X^*.$$

Lemma 8.15. *If $x^{(n)} \to x$, then $x^{(n)} \rightharpoonup x$.*

Proof From the definition of $\|\cdot\|_{X^*}$ as a supremum it follows that, for any $f \in X^*$,

$$|f(x^{(n)}) - f(x)| \leq \|f\|_{X^*}\|x^{(n)} - x\|_X.$$

Since $\|x^{(n)} - x\|_X \to 0 \quad \text{as} \quad n \to \infty$ it follows that $|f(x^{(n)}) - f(x)| \to 0 \quad \text{as} \quad n \to \infty$ and the result is proved. □

Remark 8.16. *The converse of Lemma 8.15 is not true: see Exercise 8.11.*

Lemma 8.17. *If $x^{(n)} \rightharpoonup x$, then*

$$\|x\| \leq \liminf_{n\to\infty} \|x^{(n)}\|.$$

Proof Choose $f \in X^*$ with $\|f\|_{X^*} = 1$ such that $f(x) = \|x\|$. Then

$$\|x\| = f(x) = \lim_{n\to\infty} f(x^{(n)}),$$

so

$$\|x\| \leq \liminf_{n\to\infty} |f(x^{(n)})| \leq \liminf_{n\to\infty} \|f\|_{X^*}\|x^{(n)}\|;$$

the result follows since $\|f\|_{X^*} = 1$. □

Lemma 8.18. *Weak limits are unique.*

Proof Suppose that $x^{(n)} \rightharpoonup x$ and $x^{(n)} \rightharpoonup y$. Then for any $f \in X^*$, $f(x) = \lim_{n\to\infty} f(x^{(n)}) = f(y)$. So, by Corollary 8.13, $x = y$. □

Lemma 8.19. *Weakly convergent sequence $x^{(n)} \rightharpoonup x$ in X is bounded in X:* $\sup_{n\in\mathbb{N}} \|x^{(n)}\| < \infty$.

8.3 Density and Separability

Consider a normed vector space $(X, \|\cdot\|)$ and a subset $A \subset X$.

Definition 8.20. *We say that A is* dense *in X if given $v \in X$, for every $\varepsilon > 0$, there is a $a \in A$ such that $\|v - a\| < \varepsilon$. If a space has a countable dense subset, it is called* separable *(see Figure 8.1).*

Remark 8.21. *If X is a Banach space and A is dense in X, then $X = \bar{A}$.*

Theorem 8.22. *The Banach space ℓ^p is separable for all $1 \leq p < \infty$; the Banach space ℓ^∞ is not separable.*

Proof We start by considering the case $1 \leq p < \infty$. Let $\mathbb{Q}^\sigma$ denote the collection of sequences with only a finite number of non-zero terms each of which is a rational. It follows that $\mathbb{Q}^\sigma$ is countable as it is a countable union of countable sets (those with less than or equal to n terms). It is also dense in ℓ^p: given any $v \in \ell^p$ and $\varepsilon > 0$, we may find an N such that

$$\sum_{j=N+1}^{\infty} |v_j|^p < \frac{1}{2}\varepsilon^p$$

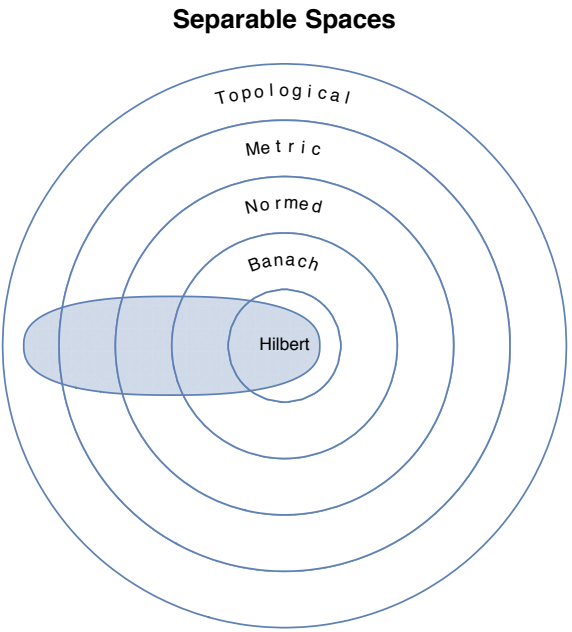


Figure 8.1 Separable spaces, as viewed as subsets of different classes of spaces.

and we may choose $w \in \mathbb{Q}^\sigma$ so that $w_j = 0$, $j \geq N+1$ and $|v_j - w_j|^p < \frac{1}{2N}\varepsilon^p$. Then

$$\|v - w\|^p = \sum_{j=1}^{N} |v_j - w_j|^p + \sum_{j=N+1}^{\infty} |v_j|^p < \varepsilon^p.$$

It follows that $\|v - w\|_{\ell^p} < \varepsilon$ as required.

Now consider the case $p = \infty$. Suppose that A is dense in ℓ^∞. Note that the uncountable set $L = \{0,1\}^\infty$ is a subset of ℓ^∞. We show that for every element of L there exists an element of A. Fix an $\varepsilon < \frac{1}{2}$. If $v, w \in L$ and $v \neq w$, then there is index j such that $v_j \neq w_j$. Without loss of generality we may take $v_j = 1$ and $w_j = 0$. By the density of A we deduce the existence of $a, b \in A$ such that $\|a - v\|_{\ell^\infty} < \varepsilon$ and $\|b - w\|_{\ell^\infty} < \varepsilon$. Hence $a_j > \frac{1}{2}$ and $b_j < \frac{1}{2}$ so that $a_j \neq b_j$. Thus A is uncountable and hence ℓ^∞ cannot be separable as A was an arbitrary dense set. □

Example 8.23.

- $L^p(D;\mathbb{R}^m)$ is separable for any $p \in [1,\infty)$.

- $L^\infty(D;\mathbb{R}^m)$ is not separable.

◊

8.4 Completion and Sobolev Space H_0^1

Definition 8.24. *Let $\left(X, \|\cdot\|\right)$ be a Banach space. Then, for a subset $Y \subset X$, the* completion *of Y in X is*

$$\overline{Y} \coloneqq \left\{ u \in X \;\middle|\; u = \lim_{n\to\infty} u^{(n)},\ \forall \{u^{(n)}\} \subset Y \right\}.$$

Remark 8.25. *The completion of a vector subspace Y of elements in Banach space $\left(X, \|\cdot\|\right)$ is itself a Banach space with respect to the same norm as that on X, by construction. If the subspace Y is dense in X then, by definition, the completion of Y in X is X itself.*

Definition 8.26. *Let $D \subset \mathbb{R}^d$ be bounded and open. Then $H_0^k(D;\mathbb{R})$ is the completion of $C_c^\infty(\overline{D};\mathbb{R})$ in $H^k(D;\mathbb{R})$.*

Let us now consider the case $k = 1$ while noting that similar results can be obtained for any k. By definition, H_0^1 is a Banach space and hence a Hilbert space with the H^1 inner product. Furthermore, the following theorem holds.

Theorem 8.27. *Let $D \subset \mathbb{R}^d$ be bounded and open. The space $H_0^1(D;\mathbb{R})$ is a Hilbert space with inner product*

$$\langle u, v\rangle_{H_0^1} \coloneqq \langle \nabla u, \nabla v\rangle_{L^2},$$

and corresponding norm

$$\|u\|_{H_0^1} = \|\nabla u\|_{L^2}.$$

Moreover, the H^1 and H_0^1 norms are equivalent.

One side of the proof is trivial. The other side relies on the *Poincaré inequality* which is given here without proof.

Lemma 8.28 (Poincaré inequality)**.** *Let $D \subset \mathbb{R}^d$ be bounded and open. For any $u \in H_0^1(D;\mathbb{R})$, there exists a $C_p > 0$ independent of u such that*

$$\|u\|_{L^2} \le C_p \|\nabla u\|_{L^2}.$$

8.5 Bases

Definition 8.29. *Let V be a vector space over $\mathbb{K}$. Let $E = \{e_\alpha\}_{\alpha\in I} \subseteq V$ be a linearly independent subset of elements such that each $v \in V$ can be written uniquely in the form*

$$v = \sum_{j=1}^{k} v_j e_{\alpha_j}$$

for some finite subset of indices $\alpha_1, \dots, \alpha_k \in I$, with I an abstract index set, and scalars $v_1, \dots, v_k \in \mathbb{K}$. Then E is called a Hamel basis *for V.*

Theorem 8.30. *Every vector space V has a Hamel basis E. Moreover, if V is infinite-dimensional, then E is uncountable.*

It is useful to work with a different notion of basis that allows for infinite linear combinations:

Definition 8.31. *Let $(V, \|\cdot\|)$ be a Banach space over $\mathbb{R}$. A* Schauder basis *is a countable collection of elements $\{\varphi^{(n)}\}_{n\in\mathbb{N}}$ in V such that, for every element $v \in V$, there is a unique sequence $\{v_j\}$ in $\mathbb{R}$ such that*

$$\lim_{n\to\infty} \left\| v - \sum_{j=1}^{n} v_j \varphi_j \right\| = 0.$$

Example 8.32. The space ℓ^p, $p \in [1, \infty)$ has a Schauder basis; ℓ^∞, which is not separable, does not. ◊

Remark 8.33. *The existence of a Schauder basis implies separability of a space. However, the converse is not true.*

Exercises

8.1 Does there exist a Banach space $(X, \|\cdot\|)$ such that X is isometrically isomorphic to X^*, but X is not Hilbert (i.e., the norm $\|\cdot\|$ does not arise from an inner product)? Give an example or prove otherwise.

8.2 Let c_0 denote the subspace of ℓ^∞ comprising sequences which converge to zero and equipped with the ℓ^∞ norm. Show that this space is separable.

8.3 Prove that a Banach space with a Schauder basis is necessarily separable.

8.4 Let $(X, \|\cdot\|)$ be an infinite-dimensional Banach space. Choose a linearly independent set $\{e_i\}_{i\in\mathbb{N}} \subseteq X$ such that $\|e_i\| = 1$ for each i. Then there exists a Hamel basis B for X such that $\{e_i\}_{i\in\mathbb{N}} \subseteq B$.

(a) Define $\varphi : X \to \mathbb{R}$ by $\varphi(e_i) = i$, and $\varphi(b) = 0$ for all $b \in B \setminus \{e_i\}_{i\in\mathbb{N}}$. Extend to the whole of X by linearity. Show that this defines an unbounded linear functional on X.

(b) Define the map $S : X \to X$ by $Su = u - 2\varphi(u)e_1$. Show that $S^2 = I$.

(c) Define the norm $\|\cdot\|_*$ on X by $\|u\|_* = \|Su\|$. Show that $\|\cdot\|_*$ and $\|\cdot\|$ are not equivalent, but the space $(X, \|\cdot\|_*)$ is complete.

8.5 Let $(X, \|\cdot\|)$ be a Banach space and let $S \subseteq X$. We will say that $u \in S$ is an interior point of S if there exists $\varepsilon > 0$ such that $B_\varepsilon(u) \subseteq S$, and we will denote $\mathrm{int}(S)$ the set of interior points of S. Show that the only subspace M of X with $\mathrm{int}(M) \neq \varnothing$ is $M = X$.

8.6 Show that any finite-dimensional subspace M of a Banach space $(X, \|\cdot\|)$ is closed.

8.7 Let $(X, \|\cdot\|)$ be an infinite-dimensional Banach space and let $S \subseteq X$. Denote by $\overline{S}$ the closure of S in X, that is, the union of S and its limit points. We will say that S is nowhere dense if $\mathrm{int}(\overline{S}) = \varnothing$. The Baire category theorem states that the complete space X cannot be the countable union of nowhere dense sets. Using Ex. 8.5, 8.6, show that X cannot be the countable union of finite-dimensional subspaces.

8.8 Let $\ell^0 \subseteq \mathbb{R}^\infty$ be the set of sequences with finitely many non-zero entries. Using Ex. 8.7, show that there does not exist a choice of norm $\|\cdot\|$ such that $(\ell^0, \|\cdot\|)$ is complete.

8.9 Let $(X, \|\cdot\|)$ be an infinite-dimensional Banach space. Using Ex. 8.7, show that there cannot exist a countable Hamel basis for X.

8.10 Show that the space $BC^0(\mathbb{R})$ of bounded continuous functions on $\mathbb{R}$, equipped with the supremum norm, is not separable.

8.11 Let $\ell^2 = \ell^2(\mathbb{Z}^+; \mathbb{R})$. Show that, if $v \in \ell^2$, then $v_j \to 0$ as $j \to \infty$. Now consider the seqeunce $\{v^{(n)}\}$ where each $v^{(n)} \in \ell^2$ is defined by $v_j^{(n)} = \delta_{jn}$. Show that sequence $\{v^{(n)}\}$ converges weakly to 0 but does not converge to 0, as $n \to \infty$.

9
Continuous Embedding

Continuous embedding is concerned with addressing the following question: given a vector space equipped with two different norms, can one determine whether the set of elements finite under one norm is a subset of the set of elements finite under the other norm?

9.1 Motivating Discussion

By Theorem 4.30 we know that all norms are equivalent in finite dimensions, and on $\mathbb{R}^n$ in particular. This implies that the set of elements of $\mathbb{R}^n$ which are finite under any norm is exactly the same. To see that the situation is more complicated in the infinite-dimensional setting, we introduce the space of countably infinite real-valued sequences

$$F = \{f : \mathbb{N} \to \mathbb{R}\}$$

which may be thought of as $\mathbb{R}^\infty$. We equip F with the norm

$$\|f\|_{\ell^p} = \left(\sum_{j\in\mathbb{N}} |f_j|^p\right)^{\frac{1}{p}},$$

for any $p \in [1, \infty)$. By ℓ^p we denote the subset of elements of F for which this norm is finite.

Example 9.1. Consider now the sequence $f_j = j^{-r}$ for some $r < 1$. Then

$$\|f\|_{\ell^p} = \left(\sum_{j\in\mathbb{N}} j^{-pr}\right)^{\frac{1}{p}}.$$

This is finite only for $p > 1/r$. This demonstrates that ℓ^p will contain different sets of sequences if p varies. ◊

Recalling the generalization of the ℓ^p spaces to include the case $p = \infty$, we have the following theorem:

Theorem 9.2. *For every $1 \leq r \leq s \leq \infty$ it follows that $\|u\|_{\ell^s} \leq \|u\|_{\ell^r}$.*

Proof To prove this, first note that

$$\max_{1 \leq j \leq J} |u_j| \leq \left(\sum_{j=1}^{\infty} |u_j|^r \right)^{\frac{1}{r}},$$

and thus

$$\sup_j |u_j| = \|u\|_{\ell^\infty} \leq \left(\sum_{j=1}^{\infty} |u_j|^r) \right)^{\frac{1}{r}}$$

which establishes the case $s = \infty$. For the case $s < \infty$, let $u \in \ell^r$ and define the sequence v by $v_j = u_j / \|u\|_{\ell^r}$. Note that

$$\sup_j |v_j| = \|v\|_{\ell^\infty} = \frac{\|u\|_{\ell^\infty}}{\|u\|_{\ell^r}} \leq 1.$$

Then, since $|v_j| \leq 1$ for all j and since $s \geq r$,

$$\|v\|_{\ell^s}^s = \sum_{j=1}^{\infty} |v_j|^s = \sum_{j=1}^{\infty} |v_j|^r |v_j|^{s-r} \leq \sum_{j=1}^{\infty} |v_j|^r = \|v\|_{\ell^r}^r = 1.$$

Therefore, $\|v\|_{\ell^s}^s \leq 1$ and hence $\|u\|_{\ell^s}^s \leq \|u\|_{\ell^r}^s$. Taking the s-root of both sides concludes the proof. □

9.2 General Setting

Let $\left(X, \|\cdot\|_X\right)$ and $\left(Y, \|\cdot\|_Y\right)$ denote two normed vector spaces with the property that $X \subseteq Y$. We define the *inclusion map* $\iota \colon X \hookrightarrow Y$ by $\iota u = u$ for every element $u \in X$. In other words, this map takes an element of X, which is a subset of Y, and maps it to exactly the same element in Y, now viewed as an element of that bigger space.

Definition 9.3. *X is said to be* continuously embedded *in Y if there is a constant $C > 0$ such that, for all $u \in X$, $\|u\|_Y \leq C\|u\|_X$.*

Implicit in this definition is that both X and Y comprise elements of a larger common vector space and are defined to be the subset of elements

of that larger vector space which are finite in their respective norms. It thus follows that if $\|u\|_Y \leq C\|u\|_X$ for all $u \in X$, then $X \subseteq Y$.

Lemma 9.4. *X is continuously embedded in Y if and only if the inclusion map ι is a bounded linear operator.*

Proof The proof is a simple consequence of definitions of the norm and continuous embedding:

$$\|\iota\|_{\mathcal{L}(X,Y)} = \sup_{x \in X \setminus \{0\}} \frac{\|\iota x\|_Y}{\|x\|_X},$$

with the last expression being bounded if and only if X is continuously embedded in Y. □

If X is continuously embedded in Y, and Y is continuously embedded in X, then the norms $\|\cdot\|_X, \|\cdot\|_Y$ are equivalent; in this situation the spaces X and Y are called *equivalent spaces* since they are essentially the same: they contain precisely the same elements. However, unlike in the finite-dimensional situation, not all norms are equivalent.

Example 9.5. Let $X = \left(\mathbb{R}^n, \|\cdot\|_X\right)$ and $Y = \left(\mathbb{R}^n, \|\cdot\|_Y\right)$. Then, by norm equivalence in finite dimensions (Theorem 4.30), X is continuously embedded in Y and vice versa: both X and Y contain exactly the same set of elements. ◊

Example 9.6. The space ℓ^r is continuously embedded into ℓ^s for every $1 \leq r \leq s \leq \infty$ and, in fact, $\|u\|_{\ell^s} \leq \|u\|_{\ell^r}$. This follows from Theorem 9.2. ◊

9.3 Sequences – Functions Defined on $\mathbb{N}$

Definition 9.7. *Let $\{\omega_j\}_{j=1}^{\infty}$ denote a strictly positive sequence of real numbers and define, for functions $f : \mathbb{N} \to \mathbb{R}$,*

$$\|f\|_{\ell^p_\omega} := \left(\sum_{j=1}^{\infty} \omega_j |f_j|^p\right)^{1/p},$$

where $p \in [1, \infty)$. Define ℓ^p_ω to be the space of all sequences for which the norm $\|\cdot\|_{\ell^p_\omega}$ is finite:

$$\ell^p_\omega = \ell^p_\omega(\mathbb{N}; \mathbb{R}) = \left\{ f : \mathbb{N} \to \mathbb{R} \;\middle|\; \|f\|_{\ell^p_\omega} < \infty \right\}.$$

Example 9.8. If $\omega_j = a + (-1)^j b$, with $a > b > 0$, then ℓ^p_ω and ℓ^p are equivalent:

$$(a-b)^{1/p}\|u\|_{\ell^p} \le \|u\|_{\ell^p_\omega} \le (a+b)^{1/p}\|u\|_{\ell^p}.$$

However, if $\omega_j = j^s$, $s > 1$, then the spaces are not equivalent. To see this consider the sequence $v : \mathbb{N} \to \mathbb{R}$ with

$$v_j^p = j^{-s-\epsilon}, \quad \epsilon \in (0,1).$$

Then this sequence is in ℓ^p but not in ℓ^p_ω:

$$\sum_{j=1}^{\infty} v_j^p < \infty, \quad \sum_{j=1}^{\infty} j^s v_j^p = \infty.$$

Note, however, that $\|f\|_{\ell^p} \le \|f\|_{\ell^p_\omega}$, so ℓ^p_ω is continuously embedded in ℓ^p, but not vice versa. ◊

Definition 9.9. *For any $s \in \mathbb{R}$, define the weighted ℓ^2 space $\mathcal{X}^s$ as follows:*

$$\mathcal{X}^s = \left\{ v : \mathbb{N} \to \mathbb{R} \;\middle|\; \sum_{j=1}^{\infty} j^{2s}|v_j|^2 < \infty \right\}. \tag{9.1}$$

This is a Hilbert space with the inner product

$$\langle v, w\rangle_{\mathcal{X}^s} = \sum_{j=1}^{\infty} j^{2s} v_j w_j \tag{9.2}$$

and norm

$$\|v\|^2_{\mathcal{X}^s} = \sum_{j=1}^{\infty} j^{2s}|v_j|^2. \tag{9.3}$$

Lemma 9.10. *$\mathcal{X}^1$ is continuously embedded into ℓ^r for all $r \in [1, \infty]$.*

Proof The case $r = 2$ is automatic:

$$\|u\|^2_{\ell^2} = \sum_{j=1}^{\infty} |u_j|^2 \le \sum_{j=1}^{\infty} j^2|u_j|^2 = \|u\|^2_{\mathcal{X}^1}.$$

The case $r > 2$ follows from the previous example and $r = 2$ case:

$$\|u\|_{\ell^r} \le \|u\|_{\ell^2} \le \|u\|_{\mathcal{X}^1}.$$

For the case $r \in [1,2)$, note that

$$\|u\|_{\ell^r}^r = \sum_{j=1}^{\infty} |u_j|^r = \sum_{j=1}^{\infty} j^r |u_j|^r \cdot j^{-r}.$$

We use the Hölder inequality with $p = 2/r$, $q = 2/(2-r)$, giving

$$\begin{aligned}\|u\|_{\ell^r}^r &\le \left(\sum_{j=1}^{\infty} \left(j^r |u_j|^r\right)^p\right)^{\frac{1}{p}} \left(\sum_{j=1}^{\infty} \left(j^{-r}\right)^q\right)^{\frac{1}{q}} \\ &\le \left(\sum_{j=1}^{\infty} j^2 |u_j|^2\right)^{\frac{r}{2}} \left(\sum_{j=1}^{\infty} j^{-\frac{2r}{2-r}}\right)^{\frac{2-r}{2}}\end{aligned}$$

and so

$$\|u\|_{\ell^r} \le \|u\|_{\mathcal{X}^1} \left(\sum_{j=1}^{\infty} j^{-\frac{2r}{2-r}}\right)^{\frac{2-r}{2r}}.$$

The sum is finite whenever $2r/(2-r) > 1$ and $r < 2$, or equivalently for $2/3 < r < 2$ and in particular for all $r \in [1,2)$ as required.

To see the embedding for the case $r = \infty$, note if $u \in \mathcal{X}^1$ then necessarily $u_j \to 0$, since the sequence $\{j^2|u_j|^2\}$ is summable. It follows that there exists a finite $j_* \ge 1$ such that $\|u\|_{\ell^\infty} = |u_{j_*}|$. We then see that

$$\|u\|_{\mathcal{X}^1}^2 = \sum_{j=1}^{\infty} j^2 |u_j|^2 \ge j_*^2 |u_{j_*}|^2 \ge |u_{j_*}|^2 = \|u\|_{\ell^\infty}^2.$$

□

9.4 Functions Defined on Subsets of $\mathbb{R}^d$

We start this section with a theorem that addresses the central question of this lecture in the context of certain L^p and Sobolev spaces. For this we will use properties of Fourier transform as discussed in Lecture 6.

Theorem 9.11.

(i) Let $u \in H^k(\mathbb{R}^d;\mathbb{R})$, $k > d/2$. Then there exists $C = C(k,d) > 0$ such that

$$\|u\|_{L^\infty} \le C\|u\|_{H^k}.$$

(ii) Let $u \in H^k(\mathbb{R}^d;\mathbb{R})$, $k \le d/2$. Then there exists $C = C(k,d) > 0$ such that

$$\|u\|_{L^p} \le C\|u\|_{H^k}$$

for any $p \in [2, \frac{2d}{d-2k})$, i.e., $u \in L^p(\mathbb{R}^d;\mathbb{R})$.

Proof (i) First we establish the bound

$$\int_{\mathbb{R}^d} \frac{1}{(1+|\xi|^2)^k}\,\mathrm{d}\xi < \infty \quad \text{if and only if } k > \frac{d}{2},$$

where $|\cdot|$ denotes the usual Euclidean norm. Observe that the integrand only depends on the distance from the origin, i.e., it is spherically symmetric. By changing to polar coordinates we have:

$$\begin{aligned}\int_{\mathbb{R}^d} \frac{1}{(1+|\xi|^2)^k}\,\mathrm{d}\xi &= |B_1(0)| \int_0^\infty \frac{r^{d-1}}{(1+r^2)^k}\,\mathrm{d}r \\ &\le |B_1(0)| \left(\int_0^1 \mathrm{d}r + \int_1^\infty \frac{r^{d-1}}{(2r^2)^k}\,\mathrm{d}r \right),\end{aligned}$$

where $|B_1(0)|$ denotes the volume of a unit ball in $\mathbb{R}^d$. The second integral clearly converges if and only if $k > d/2$, establishing the desired fact.

Now we prove the statement of the theorem, using this bound. We actually prove the assertion for $u \in S(\mathbb{R}^d;\mathbb{R}) \cap H^k(\mathbb{R}^d;\mathbb{R})$ (recall the definition of Schwarz space from Lecture 6). The desired result follows from the fact that the Schwarz space is dense in H^k. We write

$$\begin{aligned}\|u\|_{L^\infty} &= \|F^{-1}\hat{u}\|_{L^\infty} = \sup_{x\in\mathbb{R}^d} (2\pi)^{-d/2} \left| \int_{\mathbb{R}^d} \mathrm{e}^{i\langle \xi, x\rangle} \hat{u}(\xi)\,\mathrm{d}\xi \right| \\ &\le (2\pi)^{-d/2} \int_{\mathbb{R}^d} |\hat{u}(\xi)|\,\mathrm{d}\xi \\ &= (2\pi)^{-d/2} \int_{\mathbb{R}^d} \frac{1}{(1+|\xi|^2)^{k/2}} (1+|\xi|^2)^{k/2} |\hat{u}(\xi)|\,\mathrm{d}\xi \\ &\le (2\pi)^{-d/2} \left(\int_{\mathbb{R}^d} \frac{\mathrm{d}\xi}{(1+|\xi|^2)^k} \right)^{1/2} \left(\int_{\mathbb{R}^d} (1+|\xi|^2)^k |\hat{u}(\xi)|^2\,\mathrm{d}\xi \right)^{1/2} \\ &\le C\|u\|_{H^k(\mathbb{R}^d;\mathbb{R})},\end{aligned}$$

where in the second to last step we used Cauchy–Schwarz inequality, and in the last step we used Lemma 6.25 together with the integrability of $(1+|\xi|^2)^k$ for $k > d/2$, established at the start of the proof.

(ii) For the second part of the theorem, the idea of the proof is analogous

to the first one, with the only difference being the use of Hölder's inequality instead of Cauchy–Schwarz. In particular, for $q < 2$,

$$\left(\frac{2}{q}\right)^{-1} + \left(\frac{2}{2-q}\right)^{-1} = 1.$$

By Proposition 6.24, and taking p to be conjugate to $q \in [1, 2)$,

$$\|u\|_{L^p}^q \leq C_1 \|\hat{u}\|_{L^q}^q = C_1 \int_{\mathbb{R}^d} |\hat{u}(\xi)|^q \, \mathrm{d}\xi.$$

Then we can write

$$\begin{aligned}
\|u\|_{L^p}^q &\leq C_2 \int_{\mathbb{R}^d} \frac{1}{(1+|\xi|^2)^{kq/2}} (1+|\xi|^2)^{kq/2} |\hat{u}(\xi)|^q \, \mathrm{d}\xi \\
&\leq C_3 \left(\int_{\mathbb{R}^d} \frac{1}{(1+|\xi|^2)^{\frac{kq}{2}\frac{2}{2-q}}} \right)^{\frac{2-q}{2}} \cdot \left(\int_{\mathbb{R}^d} (1+|\xi|^2)^k |\hat{u}(\xi)|^2 \right)^{\frac{q}{2}} \\
&= C_4 \|u\|_{H^k}^q,
\end{aligned}$$

provided that

$$\frac{2kq}{2-q} > d \iff \frac{2k}{1-2/p} > d \iff p < \frac{2d}{d-2k}.$$

□

Remark 9.12. *Theorem 9.11 shows that $H^k(\mathbb{R}^d)$ is continuously embedded in $L^\infty(\mathbb{R}^d)$ if $k > d/2$ and that $H^k(\mathbb{R}^d)$ is continuously embedded in $L^p(\mathbb{R}^d)$ for $p \in [2, \frac{2d}{d-2k})$ for $k \leq d/2$.*

Lemma 9.13. *Let $D \subset \mathbb{R}^d$ be a bounded open subset of $\mathbb{R}^d$. Let $L^p = L^p(D; \mathbb{R})$. Then, L^r is continuously embedded into L^s for every $1 \leq s \leq r \leq \infty$.*

Proof We will first show the $r = \infty$ case:

$$\|f\|_{L^s}^s = \int_D |f(x)|^s \mathrm{d}x \leq \|f\|_{L^\infty}^s \int_D 1 \mathrm{d}x = |D| \, \|f\|_{L^\infty}^s.$$

Here $|D|$ is the d-dimensional volume of D, which is finite by supposition. Now let $1 \leq s \leq r < \infty$, and $p^{-1} + q^{-1} = 1$, we then have:

$$\|f\|_{L^s}^s = \int_D |f(x)|^s \mathrm{d}x \leq \left(\int_D |f(x)|^{sp} \mathrm{d}x \right)^{\frac{1}{p}} \left(\int_D 1 \mathrm{d}x \right)^{\frac{1}{q}}.$$

Choose $p = r/s$, then $q^{-1} = 1 - s/r = (r-s)/r$, and $q = r/(r-s)$.

Rewriting the previous bound we obtain:

$$\|f\|_{L^s}^s \le \left(\int_D |f(x)|^r \mathrm{d}x\right)^{\frac{s}{r}} \left(\int_D 1 \mathrm{d}x\right)^{1-\frac{s}{r}}.$$

Taking the s-root of both sides concludes the proof:

$$\|f\|_{L^s} \le \|f\|_{L^r} |D|^{\frac{1}{s}-\frac{1}{r}}.$$

□

Example 9.14. Consider functions $f : D \to \mathbb{R}^m$, $D \subseteq \mathbb{R}^d$, and let $\|\cdot\|_a$, $\|\cdot\|_b$ be two norms on $\mathbb{R}^m$. Define

$$\|f\|_X^p := \int_D \|f(x)\|_a^p \mathrm{d}x, \quad \|f\|_Y^p := \int_D \|f(x)\|_b^p \mathrm{d}x.$$

Finally, let $X = \{f : \|f\|_X < \infty\}$, $Y = \{f : \|f\|_Y < \infty\}$. Since all finite-dimensional norms are equivalent, we have the following relations for some constants $C_-, C_+ \in (0, \infty)$:

$$C_-\|u\|_b \le \|u\|_a \le C_+\|u\|_b \quad \forall u \in \mathbb{R}^m.$$

Thus,

$$\|f\|_X^p \le \int_D C_+^p \|f(x)\|_b^p \mathrm{d}x = C_+^p \|f\|_Y^p,$$
$$\|f\|_X^p \ge \int_D C_-^p \|f(x)\|_b^p \mathrm{d}x = C_-^p \|f\|_Y^p.$$

Hence, X, Y are equivalent and contain the same elements, and maps $\iota : X \to Y$ and $\iota : Y \to X$ are bounded linear operators. ◊

Exercises

9.1 Prove that the space $\mathcal{X}^1$ defined by (9.1)–(9.3) is a Hilbert space.

9.2 Let ω_j be the *random* sequence $\omega_j = 2 + \xi_j$ where ξ_j is itself an i. i. d. random sequence taking values ± 1 with probability $\frac{1}{2}$. Show that for any realization of this random sequence the space ℓ^2_ω is equivalent to ℓ^2.

9.3 Let $-\infty < r < s < \infty$. Show that $\mathcal{X}^s$ is continuously embedded in $\mathcal{X}^r$.

10
Compact Embedding

In finite-dimensional space $\mathbb{R}^n$ we can extract a convergent subsequence from any bounded sequence, but this fails to be true in infinite-dimensional spaces. Compactness, in various guises, is the concept required to address this issue. Note that there is not a single definition of "compact" – the concept arises in a number of inter-related definitions which we state precisely in this lecture. A key idea linking to the previous chapter is to ask the following question: if X is continuously embedded into Y, when is it the case that a bounded sequence in X has a convergent subsequence in Y? We address this question in what follows.

10.1 Compact Sets

Definition 10.1. *A subset S of a Banach space Y is said to be* (sequentially) compact in Y *if every sequence in S contains a convergent subsequence with limit in S. A subset S of a Banach space Y is said to be* relatively (sequentially) compact in Y *if every sequence in S contains a convergent subsequence with limit in Y, but not necessarily in S.*

We will drop "in Y" from this terminology only when it is clear which Banach space the subset is compact in.

Example 10.2. Let $Y = (\mathbb{R}^n, \|\cdot\|)$, then

$$S = \{u \in \mathbb{R}^n : \|u\| \leq 1\}$$

is compact while

$$S = \{u \in \mathbb{R}^n : \|u\| < 1\}$$

is relatively compact. This is the basic finite-dimensional property of bounded sets which is violated, in general, in infinite dimensions. ◊

Example 10.3. Let $I = [0, 1]$ and $Y = L^2(I; \mathbb{R})$, $X = H^1(I; \mathbb{R})$. Fix $R > 0$. Then

$$S_Y = \{u \in Y : \|u\|_Y < R\}$$

is not compact or relatively compact in Y; however,

$$S_X = \{u \in X : \|u\|_X < R\}$$

is relatively compact in Y. Although we do not prove these two assertions, they provide useful motivation for the concepts in this lecture. A related result is proved in Example 10.12, however. ◊

10.2 Compact Operators

Definition 10.4. *Let X, Y be Banach spaces. A linear operator $L : X \to Y$ is* compact *if LB is relatively compact in Y whenever B is bounded in X.*

Remark 10.5. *In the preceding definition we used*

$$LB := \{y \in Y : y = Lx, \text{for some } x \in B \subseteq X\}.$$

Proposition 10.6. *Let X, Y be Banach spaces and $L : X \to Y$ be linear and compact. Then $L \in \mathcal{L}(X, Y)$.*

Proof Assume for contradiction that L is not bounded. Then there exists a sequence $\{x^{(n)}\}$ in X with $\|x^{(n)}\|_X = 1$ and $\|Lx^{(n)}\|_Y \to \infty$. It follows that $\{Lx^{(n)}\}$ does not contain a convergent subsequence in Y. Hence LB is not relatively compact in Y for B the unit ball in X. This contradicts the fact that L is compact. □

10.3 Compact Embedding

Definition 10.7. *The Banach space X is compactly embedded in Banach space Y if $\iota : X \hookrightarrow Y$ is a compact operator.*

Example 10.8. Let $Y = L^2(I; \mathbb{R})$ and $X = H^1(I; \mathbb{R})$. Then $\iota : X \hookrightarrow Y$ is a compact operator since a bounded sequence in X has a convergent subsequence in Y. This is left without proof. ◊

Example 10.9. Let $Y = L^2(I;\mathbb{R})$ and $X = H^1(I;\mathbb{R})$. In this example we use the fact that X is compactly embedded in Y (see the preceding Example 10.8). Define $T : Y \to Y$ by

$$v(t) = (Tu)(t) := \int_0^t u(s)ds$$

for any $u \in Y$. Note that

$$\|v\|_{L^\infty} \leq \sup_{t\in[0,1]} \int_0^t |u(s)|ds \leq \int_0^1 |u(s)|ds \leq \left(\int_0^1 |u(s)|^2 ds\right)^{\frac{1}{2}} \left(\int_0^1 1ds\right)^{\frac{1}{2}}.$$

Hence,

$$\|v\|_{L^\infty} \leq \|u\|_{L^2}$$

and

$$\|v\|_{L^2}^2 = \int_0^1 |v(s)|^2 ds \leq \|v\|_{L^\infty}^2 \left(\int_0^1 1ds\right) \leq \|u\|_{L^2}^2;$$

thus $T : Y \to Y$ as claimed. But also

$$\frac{dv}{dt}(t) = u(t)$$

so that $\|dv/dt\|_{L^2}^2 = \|u\|_{L^2}^2$. Thus,

$$\|v\|_{H^1}^2 = \|v\|_{L^2}^2 + \|dv/dt\|_{L^2}^2 \leq 2\|u\|_{L^2}^2.$$

Hence $T : Y \to X$ and, in particular, if $B \subseteq Y$ is bounded by R:

$$\sup_{u\in B} \|u\|_Y \leq R$$

then

$$\sup_{v\in TB} \|v\|_X \leq \sqrt{2}R.$$

Thus TB is compact in Y since X is compactly embedded in Y. Hence T is compact when viewed as an operator from Y into itself. Note also that although T is bounded when viewed as an operator from Y into X, it is not compact as an operator between these spaces. ◊

Remark 10.10. *An equivalent definition of a compact operator $L : X \to Y$ is that if $\{x^{(n)}\}$ is a bounded sequence in X, then $\{Lx^{(n)}\}$ has a subsequence that converges in Y. In particular, Banach space X is compactly embedded in Banach space Y if $\{x^{(n)}\}$ bounded in X implies the existence of a subsequence which converges in Y.*

Let $s > 0$ and recall $\mathcal{X}^s \subset \ell^2$ defined by

$$\mathcal{X}^s = \{u \in \ell^2 : \|u\|_{\mathcal{X}^s} < \infty\},$$

where

$$\|u\|_{\mathcal{X}^s}^2 = \sum_{j=1}^{\infty} j^{2s} u_j^2.$$

Theorem 10.11. *$\mathcal{X}^s$ is compactly embedded in ℓ^2 for any $s > 0$.*

Proof Consider a sequence $\{u^{(n)}\}$ of elements bounded in $\mathcal{X}^s$. For each n, write $u^{(n)} = \{c_k^{(n)}\}_{k=1}^{\infty}$, where each $c_k^{(n)} \in \mathbb{R}$. Because the sequence is bounded, there is $M > 0$ such that

$$\sup_n \sum_{k=1}^{\infty} k^{2s} |c_k^{(n)}|^2 \le M.$$

In particular, for fixed k,

$$\sup_n |c_k^{(n)}| \le M^{\frac{1}{2}},$$

that is, $\{c_k^{(n)}\}_{n=1}^{\infty}$ is a bounded sequence in $\mathbb{R}$. Thus we may extract a convergent subsequence in $\mathbb{R}$:

$$c_1^{(n_1(j))} \to c_1^*, \quad \text{as} \quad j \to \infty.$$

We may then extract a further subsequence along which the preceding limit holds and, in addition,

$$(c_1^{(n_2(j))}, c_2^{(n_2(j))}) \to (c_1^*, c_2^*), \quad \text{as} \quad j \to \infty.$$

Recursing on this idea we find that for every $k \in \mathbb{N}$, there is a subsequence $\{n_k(j)\}$ such that

$$(c_1^{(n_k(j))}, c_2^{(n_k(j))}, \dots, c_k^{(n_k(j))}) \to (c_1^*, c_2^*, \cdots, c_k^*), \quad \text{as} \quad j \to \infty.$$

Now take the diagonal sequence $\tilde{u}^{(j)} = u^{(n_j(j))}$ and write $\tilde{c}_k^{(j)} = c_k^{(n_j(j))}$ for the corresponding elements of the sequence, noting that for each k we have $\tilde{c}_k^{(j)} \to c_k^*$ as $j \to \infty$. In this way, we define a candidate limit $u^* = (c_1^*, c_2^*, \cdots)$; it remains to show that $\tilde{u}^{(j)} \to u^*$ in ℓ^2.

We also have, for every ℓ,

$$\sum_{k=1}^{\ell} k^{2s} |c_k^*|^2 \le M$$

and hence

$$\sum_{k=1}^{\infty} k^{2s}|c_k^*|^2 \le M.$$

From this we deduce that

$$\sup_j \sum_{k=1}^{\infty} k^{2s}|\tilde{c}_k^{(j)} - c_k^*|^2 \le \sup_j \sum_{k=1}^{\infty} k^{2s} 2\left(|\tilde{c}_k^{(j)}|^2 + |c_k^*|^2\right) \le 4M.$$

Also

$$\begin{aligned}
\|\tilde{u}^{(j)} - u^*\|_{\ell^2}^2 &= \sum_{k=1}^{\infty} |\tilde{c}_k^{(j)} - c_k^*|^2 \\
&= \sum_{k=1}^{\ell} |\tilde{c}_k^{(j)} - c_k^*|^2 + \sum_{k=\ell+1}^{\infty} |\tilde{c}_k^{(j)} - c_k^*|^2 \\
&\le \sum_{k=1}^{\ell} |\tilde{c}_k^{(j)} - c_k^*|^2 + \frac{1}{\ell^{2s}} \sum_{k=\ell+1}^{\infty} k^{2s}|\tilde{c}_k^{(j)} - c_k^*|^2 \\
&\le \sum_{k=1}^{\ell} |\tilde{c}_k^{(j)} - c_k^*|^2 + \frac{1}{\ell^{2s}} \sum_{k=1}^{\infty} k^{2s}|\tilde{c}_k^{(j)} - c_k^*|^2 \\
&\le \sum_{k=1}^{\ell} |\tilde{c}_k^{(j)} - c_k^*|^2 + \frac{4M}{\ell^{2s}}.
\end{aligned}$$

Let $\varepsilon > 0$. Choose ℓ^{2s} such that $\frac{4M}{\ell^{2s}} < \frac{\varepsilon}{2}$, and then choose $J = J(\varepsilon)$ such that, for all $j \ge J$,

$$\sum_{k=1}^{\ell} |\tilde{c}_k^{(j)} - c_k^*|^2 < \frac{\varepsilon}{2}.$$

It follows that, for all $j \ge J$,

$$\|\tilde{u}^{(j)} - u^*\|_{\ell^2}^2 < \varepsilon.$$

Convergence of $\tilde{u}^{(j)}$ to u^* in ℓ^2 follows since ε is arbitrary. □

Example 10.12. The preceding with $s = 1$ shows that $H_0^1(I;\mathbb{R})$ is compactly embedded in $L^2(I;\mathbb{R})$. This follows from the fact that, for $v \in L^2(I;\mathbb{R})$,

$$v = \sum_{j=1}^{\infty} u_j \varphi_j, \tag{10.1}$$

where

$$\varphi_j(x) = \sqrt{2}\sin(j\pi x), \quad u_j = \int_0^1 v(x)\varphi_j(x)dx,$$

and from this identity it follows that

$$\|v\|_{L^2}^2 = \sum_{j=1}^{\infty} u_j^2 = \|u\|_{\ell^2}^2$$
$$\|v\|_{H_0^1}^2 = \sum_{j=1}^{\infty} j^2 u_j^2 = \|u\|_{X^1}^2.$$

The representation in Equation (10.1) follows from the Spectral Theorem of Lecture 13 and the calculation of the norms follows from orthogonality as introduced in Lecture 11. ◊

10.4 Compactness and Weak Convergence

The following facts relating compactness to weak convergence are generally useful and will in particular be useful to us when proving the Spectral Theorem and Singular Value Decompositions in Chapters 13 and 14.

Lemma 10.13. *Suppose that $L : X \to Y$ is a compact linear operator. If $x^{(n)} \rightharpoonup x$ in X, then $Lx^{(n)} \to Lx$ in Y.*

Proof First observe that $Lx^{(n)} \rightharpoonup Lx$ in Y; this follows because, if $f \in Y^*$, then $f \circ L \in X^*$ so that $x^{(n)} \rightharpoonup x$ implies, by the definition of weak convergence, that

$$f(Lx^{(n)}) \to f(Lx).$$

Now, if $Lx^{(n)} \not\to Lx$, then there is an $\varepsilon > 0$ and a subsequence (which we relabel) such that

$$\|Lx^{(n)} - Lx\| > \varepsilon \quad \forall n \in \mathbb{N}. \tag{10.2}$$

Lemma 8.19 demonstrates that $x^{(n)}$ is a bounded sequence in X. Thus, since L is compact, $\{Lx^{(n)}\}$ has a subsequence (which we relabel again) that converges to some $z \in Y$. Since $Lx^{(n)} \to z$, it follows that $Lx^{(n)} \rightharpoonup z$, by Lemma 8.15. Since weak limits are unique (Lemma 8.18), it follows that $z = Lx$. This contradicts (10.2). □

Corollary 10.14. *Suppose that X, Y are Banach spaces, with X compactly embedded in Y. If $x^{(n)} \rightharpoonup x$ in X, then $x^{(n)} \to x$ in Y.*

Proof The inclusion map $\iota : X \to Y$ is compact. □

Exercises

10.1 Let $(X, \|\cdot\|_X)$ be a separable Banach space. A set $S \subseteq X$ is said to be open if for any $u \in S$, there exists $\varepsilon > 0$ such that $B_\varepsilon(u) \subseteq S$. It is said to be compact if every open cover of S contains a finite subcover, i.e., for any family of open sets $\{A_\alpha\}$ in X,

$$S \subseteq \bigcup_\alpha A_\alpha \text{ implies that } S \subseteq \bigcup_{i=1}^{N} A_{\alpha_i}$$

for some finite subsequence $\{\alpha_i\}_{i=1}^N$.

(a) Show that $S \subseteq X$ is open if and only if $X \setminus S$ is closed.

(b) Show that $S \subseteq X$ is sequentially compact if and only if it is compact.

Hint: for one direction it may be helpful to first show that every open cover of a sequentially compact set in X contains a *countable* subcover.

(c) Let $(Y, \|\cdot\|_Y)$ be another Banach space, and let $f : X \to Y$ be continuous. Show that if $S \subseteq X$ is compact, then $f(S) \subseteq Y$ is compact.

10.2 Let $(X, \|\cdot\|)$ be a Banach space. Show that if $S \subseteq X$ is compact, then it is closed and bounded. Using Ex. 10.1(c), deduce that a continuous function on a compact set is bounded and attains its bounds.

10.3 Let $\{A_n\}_{n\in\mathbb{N}}$ be a sequence of non-empty compact sets in a Banach space $(X, \|\cdot\|)$. Suppose that $A_{n+1} \subseteq A_n$ for all n. Show that

$$\bigcap_{n=1}^{\infty} A_n \neq \emptyset.$$

Is this true if we only assume that each A_n is closed rather than compact?

10.4 Let $(X, \|\cdot\|)$ be a Banach space. Let $\theta \in (0, 1)$, and let $M \subset X$ be

a proper closed subspace of X. Show that there exists $u \in X$ such that $\|u\| = 1$ and

$$\operatorname{dist}(u, M) := \inf_{v \in M} \|u - v\| \geq \theta.$$

10.5 Let $(X, \|\cdot\|)$ be an infinite-dimensional Banach space. Using Ex. 10.4, show that the unit sphere $S = \{u \in X : \|u\| = 1\}$ is not compact. Show also that the closed unit ball $B = \{u \in X : \|u\| \leq 1\}$ is not compact.

10.6(a) For every $s \in \mathbb{R}$, define the following weighted ℓ^2 space:

$$\mathcal{X}^s = \left\{ v \colon \mathbb{N} \to \mathbb{R} \;\middle|\; \sum_{j=1}^{\infty} j^{2s} |v_j|^2 < \infty \right\}.$$

Show that this is a Hilbert space when equipped with the inner-product

$$\langle v, w \rangle_{\mathcal{X}^s} = \sum_{j=1}^{\infty} j^{2s} v_j w_j$$

and norm

$$\|v\|_{\mathcal{X}^s}^2 = \sum_{j=1}^{\infty} j^{2s} |v_j|^2.$$

(b) For which $s \geq 0$ and $r \in [1, \infty]$ is $\mathcal{X}^s$ continuously embedded into ℓ^r? (See Lemma 9.10 for a special case)

(c) For which s and r is $\mathcal{X}^s$ compactly embedded into $\mathcal{X}^r$?

10.7 Let $T \colon X \to Y$ be a linear operator between Banach spaces. Let $\{T_n\}$ be a sequence of compact operators with $\|T - T_n\|_{\mathcal{L}(X,Y)} \to 0$. Show that T is compact.

10.8 Let $(H, \langle \cdot, \cdot \rangle)$ be a separable Hilbert space. We will say that a bounded operator $A \in \mathcal{L}(H, H)$ is *Hilbert–Schmidt* if for any orthonormal basis $\{\varphi_j\}$ of H,

$$\|A\|_{HS}^2 := \sum_{j=1}^{\infty} \|A\varphi_j\|^2 < \infty. \tag{10.3}$$

(a) Show that the definition of $\|A\|_{HS}$ is independent of the choice of orthonormal basis $\{\varphi_j\}$.

(b) Show that a Hilbert–Schmidt operator is necessarily compact. **Hint**: Use Exercise 10.7.

(c) **(Hard)** Show that even if A is not bounded, it is possible for (10.3) to hold for some orthonormal basis $\{\varphi_j\}$.

10.9 Let $\ell^2 = \ell^2(\mathbb{Z}^+;\mathbb{R})$. Recall from Exercise 8.11 that if $v^* \in \ell^2$, then $v_j^* \to 0$ as $j \to \infty$.

(a) Consider the sequence $\{v^{(n)}\}$, where each $v^{(n)} \in \ell^2$ is defined by $v_j^{(n)} = \delta_{jn}$. Show that sequence $\{v^{(n)}\}$ is bounded in ℓ^2 but that there is no element $v^* \in \ell^2$ such that $v^{(n)} \to v^*$ in ℓ^2 as $n \to \infty$.

(b) Show, however, that $v^{(n)} \to 0$ in $\mathcal{X}^r$ as $n \to \infty$ for any $r < 0$. Discuss this in light of the last item in Exercise 10.6.

11
Orthogonality

In this lecture we lay the foundations for exploiting Hilbert space structure and orthogonality in particular. Throughout, we let $\left(H, \langle \cdot, \cdot \rangle, \|\cdot\|\right)$ be a Hilbert space over $\mathbb{K} = \mathbb{C}$ or $\mathbb{R}$. This use of the triple of space, inner product and norm serves to establish the notation used for each. Typically the norm will be that induced by the inner product and we will always assume this unless stated otherwise; other choices are possible. We start by defining the notion of "closest point" and then build on this to define orthogonal decomposition of the space H.

11.1 Closest Point

Lemma 11.1 (Parallelogram Rule)**.** *In a Hilbert space, the norm satisfies the* parallelogram rule*: for all $v, w \in H$,*

$$\|v + w\|^2 + \|v - w\|^2 = 2(\|v\|^2 + \|w\|^2). \tag{11.1}$$

Proof By expanding and using the definition of the norm in terms of inner product, and the properties of the inner product, we have

$$\begin{aligned}\|v + w\|^2 &= \langle v + w, v + w\rangle = \langle v, v\rangle + \langle w, v\rangle + \langle v, w\rangle + \langle w, w\rangle \\ &= \|v\|^2 + 2\,\mathrm{Re}(\langle v, w\rangle) + \|w\|^2.\end{aligned}$$

Similarly

$$\begin{aligned}\|v - w\|^2 &= \langle v - w, v - w\rangle = \langle v, v\rangle - \langle w, v\rangle - \langle v, w\rangle + \langle w, w\rangle \\ &= \|v\|^2 - 2\,\mathrm{Re}(\langle v, w\rangle) + \|w\|^2.\end{aligned}$$

Adding gives the desired result. □

Theorem 11.2 (Closest Point)**.** *Let U be a closed convex subset of a Hilbert space H and let $h \in H$. Then there is a unique $h' \in U$ such that*

$$\|h - h'\| = \inf\{\|h - a\| : a \in U\}.$$

Proof Let $\delta = \inf\{\|h - a\| : a \in U\}$. By definition of infimum we may hence find a sequence $\{h_n\}$ in U such that

$$\|h - h_n\|^2 \leq \delta^2 + n^{-1}. \tag{11.2}$$

To prove the statement we need to show that

(i) there exists a limit of $\{h_n\}$: $\lim_{n\to\infty} h_n = h' \in H$
(ii) $h' \in U$
(iii) h' is unique

(i) It suffices to show that $\{h_n\}$ is Cauchy, since H is Hilbert (hence, complete). The parallelogram rule gives

$$\|(h - h_n) + (h - h_m)\|^2 + \|(h - h_n) - (h - h_m)\|^2 = 2\big(\|h - h_n\|^2 + \|h - h_m\|^2\big).$$

Thus

$$\|2h - (h_n + h_m)\|^2 + \|h_n - h_m\|^2 < 4\delta^2 + 2(m^{-1} + n^{-1}).$$

Since U is convex, we have that $\frac{1}{2}(h_n + h_m) \in U$ and so, simply from the way δ is defined,

$$\|h - \frac{1}{2}(h_n + h_m)\|^2 \geq \delta^2.$$

Thus

$$\|h_n - h_m\|^2 < 4\delta^2 + 2(m^{-1} + n^{-1}) - 4\|h - \frac{1}{2}(h_n + h_m)\|^2 \leq 2(m^{-1} + n^{-1}).$$

It follows that $\{h_n\}$ is Cauchy in H and hence has a limit $h' \in H$.

(ii) Since U is closed it follows that $h' \in U$.

(iii) To show uniqueness, suppose for contradiction that there exists $h'' \in U$ such that $\|h - h''\| = \delta$ and $h'' \neq h'$. Then we deduce that, by convexity, $\frac{1}{2}(h' + h'') \in U$. Thus, $\|h - \frac{1}{2}(h' + h'')\| \geq \delta$. By the parallelogram rule we have

$$\|(h - h') + (h - h'')\|^2 + \|(h - h') - (h - h'')\|^2 = 2\big(\|h - h'\|^2 + \|h - h''\|^2\big) = 4\delta^2.$$

It follows that

$$\|h' - h''\|^2 = 4\delta^2 - 4\|h - \frac{1}{2}(h' + h'')\|^2 \leq 0$$

and hence that $h'' = h'$ establishing uniqueness.

□

Example 11.3. Let $z \in H$ and define

$$U = \text{span}\{z\} = \{u \in H : u = \alpha z, \alpha \in \mathbb{R}\}.$$

U is convex because it is a subspace of H. To see that it is closed, let $\{u^{(n)}\}$ be a sequence in U with limit $u \in H$. Then $u^{(n)} = \alpha_n z$ with $\{\alpha_n\}_{n\in\mathbb{N}}$ a sequence in $\mathbb{R}$. The fact that $\{u^{(n)}\}$ converges means it is Cauchy; this may be used to show that $\{\alpha_n\}$ is also Cauchy. Hence $\{\alpha_n\}_{n\in\mathbb{N}}$ converges to limit $\alpha \in \mathbb{R}$. Now note

$$\|u^{(n)} - \alpha z\| = \|(\alpha_n - \alpha)z\| = |\alpha_n - \alpha|\|z\|.$$

Because $\alpha_n \to \alpha$ in $\mathbb{R}$ it follows that $u^{(n)} \to \alpha z$ in H, and since $\alpha z \in U$, the closedness of U is established. ◊

11.2 Orthogonal Decomposition

Definition 11.4. *Two elements v, w of an inner product space H are said to be* orthogonal *if $\langle v, w\rangle = 0$; we write $v \perp w$. For a subspace U in H, we define the* orthogonal complement *by*

$$U^\perp = \{v \in H : \langle v, w\rangle = 0 \quad \forall w \in U\}.$$

Remark 11.5. *Finite-dimensional subspaces are necessarily closed. However, subspaces in infinite-dimensional spaces are not always closed.*

Definition 11.6. *A Schauder basis in H for which, additionally, $\langle \varphi_j, \varphi_k\rangle = \delta_{jk}$, is called an* orthonormal basis.

Lemma 11.7. *The orthogonal complement satisfies the following three properties:*

- *For any subspace $U \subset H$, $U^\perp$ is a closed subspace in H.*
- *$U \cap U^\perp = \{0\}$.*
- *If U is a closed subspace, then $(U^\perp)^\perp = U$.*

Proof

- Let $\{v_n\}_{n\in\mathbb{N}} \subset U^\perp$ with $v_n \to v \in H$. For any $u \in U$ we have that
$$\langle v, u\rangle = \left\langle \lim_{n\to\infty} v_n, u \right\rangle = \lim_{n\to\infty} \langle v_n, u\rangle = 0$$
since the map $v \mapsto \langle v, u\rangle$ is continuous. Hence $v \in U^\perp$, and so $U^\perp$ is closed.
- Let $v \in U \cap U^\perp$, then for each $u \in U$, $\langle v, u\rangle = 0$ since $v \in U^\perp$. In particular, since $v \in U$, we have $\langle v, v\rangle = \|v\|^2 = 0$ and so $v = 0$.
- We recall that
$$U^\perp = \{v \in H : \langle v, u\rangle = 0 \text{ for all } u \in U\}$$
$$(U^\perp)^\perp = \{w \in H : \langle w, v\rangle = 0 \text{ for all } v \in U^\perp\}.$$

  If $u \in U$, then for any $v \in U^\perp$ we have $0 = \langle v, u\rangle = \langle u, v\rangle$, and so $u \in (U^\perp)^\perp$. It follows that $U \subseteq (U^\perp)^\perp$. Now let $w \in (U^\perp)^\perp$. Given an orthonormal basis $\{e_j\}_{j=1}^\infty$ for U, define the element
$$w' = w - \sum_{j=1}^{\infty} \langle e_j, w\rangle e_j.$$
  The sum converges to an element of U since U is assumed closed. We have $\langle e_k, w'\rangle = 0$ for all k, and so $w' \in U^\perp$. But then $w' \in U^\perp \cap (U^\perp)^\perp$, and so must be zero by the previous part. It follows that $w \in U$, and so $U = (U^\perp)^\perp$.

□

Theorem 11.8. *If U is a closed subspace of an inner product space H, then the closest point $h' \in U$ to $h \in H$ satisfies $h - h' \in U^\perp$. Thus $H = U \oplus U^\perp$: any $h \in H$ can be uniquely decomposed as $h = p + q$, where $p \in U$ and $q \in U^\perp$.*

Proof We want to show that $v = h - h'$ satisfies $\langle v, w\rangle = 0$ for all $w \in U$. This would imply $v \in U^\perp$. Consider $\|h - (h' - tw)\| = \|v + tw\|$ for $t \in \mathbb{C}$ and define
$$F(t) \coloneqq \|h - (h' - tw)\|^2.$$
Note that F is minimized at $t = 0$, by definition of h'. Now,
$$\begin{aligned} F(t) &= \langle v + tw, v + tw\rangle \\ &= \|v\|^2 + \langle tw, v\rangle + \langle v, tw\rangle + \|tw\|^2 \\ &= \|v\|^2 + 2\,\mathrm{Re}\big(t\langle w, v\rangle\big) + |t|^2\|w\|^2. \end{aligned}$$

If t is real, then the fact that $F(t)$ is minimized at $t = 0$ gives $\mathrm{Re}\big(\langle w, v\rangle\big) = 0$ (as the last expression can be viewed as a quadratic equation in t). If $t = is$ with s real then the fact that $F(is)$ is minimized at $s = 0$ gives $\mathrm{Im}\big(\langle w, v\rangle\big) = 0$. Thus $\langle w, v\rangle = 0$ for all $w \in U$ and the first statement is proved.

From above we know that any $h \in H$ can be written as $h = p + q$, with p being the closest point in U and $q \in U^\perp$. It remains to show that this is unique. Assume for contradiction that $h = r + s$ with $r \in U$ and $s \in U^\perp$. We then have that $p + q = r + s$, and so $p - r = s - q$. Therefore, we can write

$$\|p - r\|^2 = \langle p - r, p - r\rangle = \langle s - q, p - r\rangle = 0,$$

where the last equality follows from the fact that $(p - r) \in U$ and $(s - q) \in U^\perp$. Thus $p = r$ and $q = s$, and uniqueness is established. □

Example 11.9. Let $\{\varphi^{(j)}\}_{j\in\mathbb{N}}$ be an orthonormal basis for H and let

$$U \coloneqq \mathrm{span}\{\varphi^{(j)}\}_{j=1}^{p'} = \Big\{h \in H : h = \sum_{j=1}^{p'} h_j\varphi^{(j)},\ \{h_j\}_{j=1}^{p'} \subset \mathbb{R}\Big\}$$

for some $p' \in \mathbb{N}$. If $h \in U$, then the $\{h_j\}_{j=1}^{p'}$ are uniquely defined by

$$\begin{aligned}\langle \varphi^{(i)}, h\rangle &= \sum_{j=1}^{p'} h_j\langle \varphi^{(i)}, \varphi^{(j)}\rangle \\ &= \sum_{j=1}^{p'} h_j\delta_{ij} \\ &= h_i\end{aligned}$$

for $i \in \{1, \dots, p'\}$. We now show that U is closed. Let $h^{(n)}$ denote a sequence in U which converges to $h \in H$ in H. Then we may write

$$h^{(n)} = \sum_{j=1}^{p'} h_j^{(n)}\varphi^{(j)}$$

and note that, since $h^{(n)}$ also converges weakly in H,

$$h_j^{(n)} = \langle \varphi^{(j)}, h^{(n)}\rangle \to h_j \coloneqq \langle \varphi^{(j)}, h\rangle$$

in $\mathbb{R}$. Thus, using orthogonality,

$$\|h^{(n)} - \sum_{j=1}^{p'} h_j \varphi^{(j)}\|^2 = \sum_{j=1}^{p'} |h_j^{(n)} - h_j|^2.$$

The convergence of the $h_j^{(n)}$ to the h_j, for each $j \in \{1, \cdots, p'\}$, shows that

$$\|h^{(n)} - \sum_{j=1}^{p'} h_j \varphi^{(j)}\|^2 \to 0$$

and hence

$$h^{(n)} \to h = \sum_{j=1}^{p'} h_j \varphi^{(j)} \in U$$

establishes that U is closed. ◊

11.3 Orthogonal Projection

Definition 11.10. *Given a closed subspace U of an inner product space H, we define the operator $P : H \to U$ as follows. For any $h \in H$, define $Ph = p$, where p is the closest point to h in U. We call P* orthogonal projection *onto U.*

Theorem 11.11. *The operator P satisfies:*

- *P is linear,*
- *$P^2 = P$,*
- *$\|P\| = 1$,*

where the operator norm is in $\mathcal{L}(H, H)$.

Proof By Theorem 11.8 we have that, for any $\alpha, \beta \in \mathbb{K}$ and $v, w \in H$, there is a unique $q \in U^\perp$ such that

$$\alpha v + \beta w = P(\alpha v + \beta w) + q;$$

furthermore, there are unique $q_1, q_2 \in U^\perp$ such that

$$v = Pv + q_1, \quad w = Pw + q_2.$$

Together these identities imply that

$$P(\alpha v + \beta w) + q = \alpha(Pv + q_1) + \beta(Pw + q_2) = (\alpha Pv + \beta Pw) + (\alpha q_1 + \beta q_2).$$

However $q, \alpha q_1, \beta q_2$ are all elements of $U^\perp$ whilst the remaining terms

in the identity are elements of U. Equating the components in U, which we may do because of the uniqueness of the decomposition given by Theorem 11.8, gives

$$P(\alpha v + \beta w) = \alpha Pv + \beta Pw$$

as required for linearity.

Let $h = p + q$ denote the unique decomposition of h into $p \in U$ and $q \in U^\perp$. Then $Ph = p$. Thus $P^2h = Pp$. However, since the closest point to p in U is p (by uniqueness), $Pp = p$. Thus $P^2h = p = Ph$. Since this is true for all elements $h \in H$, we deduce that $P^2 = P$.

Finally, since $h - p \perp p$ (recall that $(h - p) \in U^\perp$ by uniqueness),

$$\|h\|^2 = \|h - p + p\|^2 = \|h - p\|^2 + \|p\|^2 \geq \|p\|^2.$$

Therefore, $\|Ph\| \leq \|h\|$ for all $h \in H$ demonstrating that

$$\|P\| = \sup_{\|h\|=1} \|Ph\| \leq 1.$$

Furthermore, if $v \in U$, then $Pv = v$ and $\|Pv\| = \|v\|$ demonstrating that $\|P\| \geq 1$. Hence $\|P\| = 1$ as required. □

Example 11.12. Let $U = \text{span}\{z\}$ for some $z \in H$ with $\|z\| = 1$. Then, for $h \in H$,

$$p = Ph = \langle z, h\rangle z.$$

Thus $P = z \otimes z$ and $\|P\| = \|z\|^2 = 1$. A similar argument may be applied if z is not normalized to size 1 in H. ◊

11.4 Adjoints

Definition 11.13. *Let $T \in \mathcal{L}(H, H)$. The* adjoint *of T, denoted T^*, is the element of $\mathcal{L}(H, H)$ with the property that*

$$\langle u, Tv\rangle = \langle T^*u, v\rangle \quad \forall u, v \in H.$$

Then $T \in \mathcal{L}(H, H)$ is symmetric *if $T^* = T$, that is, if*

$$\langle u, Tv\rangle = \langle Tu, v\rangle \quad \forall u, v \in H.$$

Definition 11.14. *Operator $T \in \mathcal{L}(H, H)$ is said to be* normal *if $TT^* = T^*T$.*

Example 11.15. The adjoint of matrix $A \in \mathbb{R}^{n\times n}$ is $A^\top$ and matrix A is symmetric if $A^\top = A$. A symmetric matrix is normal. ◊

Remark 11.16. *Sometimes a symmetric operator $T \in \mathcal{L}(H, H)$ is termed* self-adjoint. *We, however, reserve this terminology for the setting in which $T \in \mathcal{L}(D(T), H)$, where the domain of T, $D(T)$, is dense in H but not equal to H.*

Exercises

11.1 Let $A \in \mathbb{C}^{m\times n}$. Show that A^*A is Hermitian and positive semi-definite.

11.2 Show that, if a norm on a vector space over the real numbers $\mathbb{R}$ comes from an inner product, then the inner product can be recovered from the *polarization identity*

$$4\langle v, w\rangle = \|v + w\|^2 - \|v - w\|^2. \tag{11.3}$$

11.3 Show that, if a norm on a vector space H over the real numbers $\mathbb{R}$ satisfies the parallelogram law (11.1), and if you define a map $\langle\cdot,\cdot\rangle\colon H \times H \to \mathbb{R}$ by (11.3), then the resulting map satisfies the axioms of an inner product.

11.4 Let $\{e_j\}_{j=1}^n$ be an orthonormal set of vectors in an inner product space H, so that $\langle e_i, e_j\rangle = \delta_{ij}$. Assume that U comprises the linear span of $\{e_j\}_{j=1}^n$: the set of all vectors which may be represented as a linear combination of the $\{e_j\}_{j=1}^n$. Prove that the closest point $p \in U$ to $h \in H$ is

$$p = \sum_{j=1}^{n} \langle e_j, h\rangle e_j.$$

11.5 Show that the functions $\{\varphi_n\}_{n\in\mathbb{Z}^+}$ defined by

$$\varphi_0(x) = \frac{1}{\sqrt{2}}, \quad \varphi_{2j}(x) = \cos(j\pi x), \quad \varphi_{2j-1}(x) = \sin(j\pi x)$$

are orthonormal with respect to the inner product

$$\langle u, v\rangle_{L^2([-1,1];\mathbb{R})} = \int_{-1}^{1} u(x)v(x)dx.$$

Show in addition that the functions $\{\psi_n\}_{n\in\mathbb{Z}}$ defined by

$$\psi_n(x) = \frac{1}{\sqrt{2}} \exp(in\pi x)$$

are orthonormal with respect to the inner product

$$\langle u, v\rangle_{L^2([-1,1];\mathbb{C})} = \int_{-1}^{1} \overline{u(x)} v(x) dx.$$

In both cases give formulae for an expansion of a function f in the form

$$f(x) = \sum_{n\in\mathbb{Z}^+} f_n \varphi_n(x), \quad f(x) = \sum_{n\in\mathbb{Z}} f_n \psi_n(x).$$

In the complex-valued setting, if the function is real, then the two expansions should coincide. Demonstrate the relationship between the coefficients in the two expansions.

11.6 Define

$$\psi(t) = \begin{cases} 1, & t \in [0, \frac{1}{2}) \\ -1, & t \in [\frac{1}{2}, 1) \\ 0, & t \notin [0, 1). \end{cases}$$

From this we construct the family of Haar wavelets

$$\psi_{j,n}(t) = \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j}\right)$$

defined for $(j, n) \in \mathbb{Z}^2$. Show that the Haar wavelets are orthonormal in $L^2(\mathbb{R}; \mathbb{R})$.

11.7 Let X, Y be normed vector spaces and $T \in \mathcal{L}(X, Y)$.

(a) Show that there exists a unique bounded linear operator $T' \in \mathcal{L}(Y^*, X^*)$ such that

$$(T'f)(u) = f(Tu) \quad \text{for all } u \in X, f \in Y^*.$$

(b)(i) Show that if $Y = \{0\}$, then $T = T' = 0$.

(ii) Assume that $Y \neq \{0\}$. Show that $Y^* \neq \{0\}$.

(iii) Given $u \in X$ with $\|u\|_X = 1$, show that there exists $f \in Y^*$ such that $f(Tu) = \|Tu\|_Y$.

(iv) Prove that $\|T'\|_{\mathcal{L}(Y^*,X^*)} = \|T\|_{\mathcal{L}(X,Y)}$.

(c) If X, Y are Hilbert spaces, how is $T' \in \mathcal{L}(Y^*, X^*)$ related to the adjoint operator $T^* \in \mathcal{L}(Y, X)$?

Hint: You will need to use Theorem 8.11.

11.8 Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and let P, Q be densely-defined

self-adjoint operators on H. Show that for any $\psi \in D(P) \cap D(Q) \cap D(PQ) \cap D(QP)$ with $\|\psi\| = 1$,

$$\|P\psi - \langle\psi, P\psi\rangle\psi\|\|Q\psi - \langle\psi, Q\psi\rangle\psi\| \geq \frac{1}{2}|\langle\psi, (PQ - QP)\psi\rangle|.$$

Remark. This is Heisenberg's uncertainty principle. It is often written as

$$\langle P - \langle P\rangle_\psi\rangle_\psi^{1/2}\langle Q - \langle Q\rangle_\psi\rangle_\psi^{1/2} \geq \frac{1}{2}|\langle[P, Q]\rangle_\psi|$$

with $\langle A\rangle_\psi$ representing the mean square of A with respect to the probability density $|\psi|^2$.

11.9 Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space, and let $A\colon H \to H$ be bounded. Let λ be an eigenvalue of A and $E_\lambda \subseteq H$ its corresponding eigenspace. Show that E_λ is closed. Show also that it is invariant, i.e., $A(E_\lambda) \subseteq E_\lambda$. When do we have $A(E_\lambda) = E_\lambda$?

11.10 For each $n \in \mathbb{Z}^+$, define the Hermite polynomial $H_n\colon \mathbb{R} \to \mathbb{R}$ by

$$H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n}\left(e^{-x^2}\right).$$

(a) Show that H_n is a polynomial of degree n.

(b) Show that for each $n, k \in \mathbb{Z}^+$,

$$\int_{-\infty}^{\infty} e^{-x^2} H_n(x) H_k(x)\, dx = \sqrt{\pi} 2^n n! \delta_{nk}.$$

You may use that $\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi}$.

(c) Show that for any $n \in \mathbb{N}$ and $x \in \mathbb{R}$,

$$H_{n+1}(x) = 2xH_n(x) - 2nH_{n-1}(x).$$

11.11 For each $n \in \mathbb{Z}^+$, define the function $\varphi_n\colon \mathbb{R} \to \mathbb{R}$ by

$$\varphi_n(x) = \frac{1}{\sqrt{2^n n!}}\left(\frac{m\omega}{\pi\hbar}\right)^{1/4} e^{-\frac{m\omega x^2}{2\hbar}} H_n\left(\sqrt{\frac{m\omega}{\hbar}}x\right)$$

where H_n is the Hermite polynomial defined in Exercise 11.10.

(a) Show that the family $\{\varphi_n\}$ is orthonormal in $L^2(\mathbb{R}; \mathbb{C})$.

(b) Show that each φ_n is an eigenvector of the operator $\hat{H}\colon D(\hat{H}) \to L^2(\mathbb{R}; \mathbb{C})$ given by

$$\hat{H} = -\frac{\hbar^2}{2m}\frac{\partial^2}{\partial x^2} + \frac{1}{2}m\omega^2\hat{x}^2$$

where for each $x \in \mathbb{R}$, $(\hat{x}^2\psi)(x) = x^2\psi(x)$. Calculate the corresponding eigenvalues λ_n.

(c) Assuming that the set $\{\varphi_n\}$ is a complete orthonormal system for $L^2(\mathbb{R}, \mathbb{C})$, write down the solution to the time-dependent Schrödinger equation for the quantum harmonic oscillator:

$$\frac{d\psi}{dt}(t) = -\frac{\mathrm{i}}{\hbar}\hat{H}\psi(t), \quad \psi(0) = \psi_0$$

where $\psi_0 \in D(\hat{H})$.

11.12 Let

$$H = L^2([0,1];\mathbb{R}) := \left\{ u : [0,1] \to \mathbb{R} \,\middle|\, \int_0^1 u(x)^2 dx < \infty \right\}$$

with inner product

$$\langle u, v \rangle_{L^2} = \int_0^1 u(x)v(x)dx.$$

Show that the family of functions $\{\varphi_j\}_{j\in\mathbb{N}}$ defined by $\varphi_j(x) = \sqrt{2}\sin(2\pi jx)$ are orthonormal with respect to the inner product. Show that they do not form an orthonormal basis for H.

11.13 Let U be the closed ball of radius R in $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$, centered at the origin. Fix $h \in H$ and find $h' \in U$ given by Theorem 11.2. To achieve this, let $V = \mathrm{span}\{h\}$ and write $h' = p + q$ with $p \in V$ and $q \in V^\perp$. Distinguish between the cases where $h \in U$ and $h \notin U$ in your answer.

12
Riesz Representation and Lax–Milgram

The Riesz representation theorem is the observation that the dual space of a Hilbert space is isometrically isomorphic to the Hilbert space itself. We prove this theorem, show that it has immediate application to the solution of linear equations, and state the celebrated Lax–Milgram theorem which is a widely used generalization of Riesz representation and useful for the solution of linear equations.

12.1 Riesz Representation Theorem

Let $\big(H, \langle\cdot,\cdot\rangle, \|\cdot\|\big)$ denote a Hilbert space. We start with the following observation showing that every element of the Hilbert space H may be identified with an element of the dual space H^*:

Lemma 12.1. *Given any $w \in H$, we can construct an element of $l_w \in H^*$ defined by $l_w(\cdot) = \langle w, \cdot\rangle$. Furthermore $\|l_w\|_{H^*} = \|w\|$.*

Proof Given any $w \in H$, define

$$l_w(v) = \langle w, v\rangle. \tag{12.1}$$

Then l_w is clearly linear, and

$$|l_w(v)| = |\langle w, v\rangle| \le \|w\|\|v\|$$

using the Cauchy-Schwarz inequality. It follows that $l_w \in H^*$ with

$$\|l_w\|_{H^*} = \inf\{K : |l_w(v)| \le K\|v\|\} = \sup_{v\in H\setminus\{0\}} \frac{|l_w(v)|}{\|v\|} \le \|w\|.$$

Choosing $v = w$ in (12.1) shows that

$$|l_w(w)| = \langle w, w\rangle = \|w\|^2$$

and hence that $\|l_w\|_{H^*} = \|w\|$. □

There is a converse to the preceding lemma, showing that every linear functional in H^* may be identified with an element of the original space H:

Theorem 12.2 (Riesz Representation Theorem)**.** *For every bounded linear functional f on a Hilbert space H, there exists a unique element $w \in H$ such that*

$$f(v) = \langle w, v \rangle \qquad \text{for all} \qquad v \in H; \tag{12.2}$$

furthermore, $\|w\|_H = \|f\|_{H^*}$. Thus $H^* \simeq H$.

Proof Let

$$K = \mathrm{Ker}(f) = \{u \in H : f(u) = 0\}.$$

Note that this is a closed subspace of H. It is clearly a vector subspace. To see that it is closed, note that if $u_n \in \mathrm{Ker}(f)$ and $u_n \to u$, then

$$|f(u)| = |f(u) - f(u_n)| = |f(u - u_n)| \leq \|f\|_{H^*}\|u - u_n\|;$$

this holds for all n, hence $|f(u)| = 0$ since $\|u - u_n\| \to 0$.

If $f = 0$, then $K = H$ and we set $w = 0$. Otherwise, if $f \neq 0$, then $K \neq H$; it follows that $K^\perp$ is non-empty. In fact, $K^\perp$ is a one-dimensional subspace of H: indeed, if $u, v \in K^\perp$, then u is proportional to v. To see this, note that we have, by linearity of f,

$$f(f(u)v - f(v)u) = 0. \tag{12.3}$$

Since $u, v \in K^\perp$, it follows that $f(u)v - f(v)u \in K^\perp$, while (12.3) shows that $f(u)v - f(v)u \in K$. It follows that $f(u)v - f(v)u = 0$, and so u and v are proportional.

As a consequence of this one-dimensional structure of $K^\perp$, we can choose $z \in K^\perp$ with $\|z\| = 1$ and decompose any $v \in H$ as

$$v = \alpha z + u$$

with $u \in K$. Note furthermore that, since $\langle z, u \rangle = 0$ and $\|z\|^2 = 1$, $\alpha = \langle z, v \rangle$. Thus

$$v = \langle z, v \rangle z + u.$$

Therefore, since $f(u) = 0$,

$$f(v) = \langle z, v \rangle f(z) = \langle \overline{f(z)}z, v \rangle.$$

Setting $w = \overline{f(z)}z$ we obtain (12.2).

To show that this choice of w is unique, suppose that

$$\langle w, v\rangle = \langle \hat{w}, v\rangle \qquad \text{for all} \qquad v \in H.$$

Then $\langle w - \hat{w}, v\rangle = 0$ for all $v \in H$, i.e., $w - \hat{w} \in H^{\perp} = \{0\}$. The same calculation as used in Lemma 12.1 shows the equality of the norms of w and f. Together, the result of this theorem and the preceding Lemma 12.1 shows that $H^* \simeq H$. □

Recall Definition 8.14 of weak convergence. The following is a straightforward consequence of the Riesz Representation Theorem:

Lemma 12.3. *Let $x_n \rightharpoonup x$ in Hilbert space H. This is equivalent to the statement that*

$$\langle x_n, y\rangle \to \langle x, y\rangle \qquad \textit{for all} \qquad y \in H.$$

We will find the following two theorems useful when proving the Spectral Theorem. [1]

Theorem 12.4. *Let H be a Hilbert space. Then any bounded sequence in H has a weakly convergent subsequence.*

Theorem 12.5. *Let H be a Hilbert space and $\{h_n\}_{n\in\mathbb{N}}$ an orthonormal set of vectors. Then $h_n \rightharpoonup 0$.*

12.2 Application to Solving PDEs

Recall, from Chapter 3, the second order elliptic PDE

$$\begin{aligned} -\nabla\cdot(a(x)\nabla u(x)) &= r(x), \quad x \in D \subset \mathbb{R}^d \\ u(x) &= 0, \qquad x \in \partial D. \end{aligned} \tag{E}$$

The objective is to find $u\colon D \to \mathbb{R}$, given $a, r\colon D \to \mathbb{R}$. In (wPDE) we introduced the weak formulation of the PDE, namely to find $u \in H_0^1(D;\mathbb{R})$ such that

$$\int_D a(x)\langle \nabla u(x), \nabla v(x)\rangle_{\mathbb{R}^d}\, dx = \int_D r(x)v(x)dx, \quad \forall v \in H_0^1(D;\mathbb{R}). \tag{WE}$$

[1] The proofs of Theorems 12.4 and 12.5 are beyond the scope of these notes. However we observe that Theorem 12.4 is a special case and consequence of a more general result concerning reflexive Banach spaces, due to Eberlein (1947). The result states that the closed unit ball is weakly sequentially compact in Banach space X if and only if X is reflexive. The notion of weakly sequentially compact generalizes Definition 10.1 to the extraction of weakly convergent subsequences. Defining reflexivity on Banach space requires the notion of weak-* convergence; all Hilbert spaces are reflexive, essentially as a consequence of the Riesz Representation Theorem.

For smooth enough a and r, a solution of (E) is also a solution of (WE):

Lemma 12.6. *Assume that $a \in C^1(D;\mathbb{R})$ and $r \in C(D;\mathbb{R})$. If $u \in C^2(D;\mathbb{R})$ solves* (E)*, then u solves* (WE)*.*

However, in practical applications it is often desirable to solve the problem under weaker assumptions on a and r; the weak formulation is useful for this purpose. Furthermore, the weak formulation is desirable for the conception of computational methods. Thus we proceed to study the weak formulation, under more relaxed assumptions on a and r. To this end we define $H = H_0^1(D;\mathbb{R})$ and recall the inner product

$$\langle u, v\rangle = \int_D \langle \nabla u(x), \nabla v(x)\rangle_{\mathbb{R}^d}\, dx.$$

We denote the induced norm by $\|\cdot\|$. We define B and f as follows:

$$B(u,v) = \int_D a(x)\langle \nabla u(x), \nabla v(x)\rangle_{\mathbb{R}^d}\, dx$$
$$f(v) = \int_D r(x)v(x)dx.$$

Assumptions 12.7. *The coefficient and source functions are such that $a \in L^\infty(D;\mathbb{R})$, $r \in L^2(D;\mathbb{R})$ and there exists $a^- \in (0,\infty)$ such that*

$$\operatorname*{ess\,inf}_{x\in D} a(x) = a^-.$$

We have the following properties of B and f:

Lemma 12.8. *Let Assumptions 12.7 hold. Then $B : H\times H \to \mathbb{R}$ is (respectively) bounded, coercive and symmetric:*

- *$\exists \alpha > 0$ such that $|B(u,v)| \le \alpha\|u\|\|v\|$;*
- *$\exists \beta > 0$ such that $B(u,u) \ge \beta\|u\|^2$;*
- *$B(u,v) = B(v,u)$ for all $u, v \in H$.*

Furthermore, $f \in \mathcal{L}(H,\mathbb{R}) = H^*$.

Proof Since $a \in L^\infty(D;\mathbb{R})$ we have, for some $a^+ \in (0,\infty)$,

$$\operatorname*{ess\,sup}_{x\in D} |a(x)| = a^+.$$

Thus

$$|B(u,v)| \le a^+|\langle u, v\rangle| \le a^+\|u\|\|v\|.$$

Similarly

$$B(u,u) \geq a^- \langle u,u\rangle = a^- \|u\|^2.$$

Symmetry is immediate from the definition. Finally, we note that by Cauchy-Schwarz,

$$|f(v)| \leq \|r\|_{L^2}\|v\|_{L^2}.$$

By the Poincaré Lemma 8.28 we deduce that

$$|f(v)| \leq C_p\|r\|_{L^2}\|v\|$$

and hence that

$$\|f\|_{H^*} = \sup_{v\neq 0} \frac{|f(v)|}{\|v\|} \leq C_p\|r\|_{L^2} < \infty$$

as required. □

We now revert to the setting of a general Hilbert space $\big(H, \langle\cdot,\cdot\rangle, \|\cdot\|\big)$. We consider the abstract problem of finding $u \in H$ such that

$$B(u,v) = f(v) \quad \forall v \in H, \tag{A}$$

where $f \in H^*$ and B is bilinear. The weak formulation of the PDE, namely (WE), is an example of this problem. However, this abstract problem has a solution in quite general settings, which include the weak PDE, but is by no means limited to it.

Definition 12.9. *$B : H \times H \to \mathbb{R}$ is a* bilinear form *if*

$$B(\alpha u_1 + \beta u_2, v) = \alpha B(u_1,v) + \beta B(u_2,v), \quad \forall \alpha,\beta \in \mathbb{R},\ \forall u_1,u_2,v \in H$$

and

$$B(u, \alpha v_1 + \beta v_2) = \alpha B(u,v_1) + \beta B(u,v_2), \quad \forall \alpha,\beta \in \mathbb{R},\ \forall u,v_1,v_2 \in H.$$

Corollary 12.10. *Suppose that H is a real Hilbert space and $B : H \times H \to \mathbb{R}$ is a bilinear form satisfying the following three conditions:*

(i) bounded*: there exists $\alpha \geq 0$ such that*

$$|B(u,v)| \leq \alpha\|u\|\|v\| \qquad \textit{for all} \qquad u,v \in H;$$

(ii) coercive*: there exists $\beta > 0$ such that*

$$B(u,u) \geq \beta\|u\|^2 \qquad \textit{for all} \qquad u \in H;$$

(iii) symmetric*:*

$$B(u,v) = B(v,u) \qquad \textit{for all} \qquad u,v \in H.$$

Then for any $f \in H^*$, there exists a unique $u_f \in H$ solving (A) *and satisfying*

$$\|u_f\| \leq \beta^{-1}\|f\|_* . \tag{12.4}$$

Proof B is an alternative inner product on H; from (ii), $B(u,u) \geq 0$ with equality if and only if $u = 0$; B satisfies the linearity requirements of an inner product; and B is symmetric by assumption. Moreover, by (i) and (ii) we have

$$\beta\|u\|^2 \leq B(u,u) \leq \alpha\|u\|^2 \qquad \text{for all} \qquad u \in H,$$

and so the norm induced by the inner product $B(\cdot,\cdot)$ is equivalent to $\|\cdot\|$. It follows that $(H, B(\cdot,\cdot))$ is again complete, and hence Hilbert.

We can immediately apply the Riesz Representation Theorem to deduce the existence of a unique u_f satisfying (A). The bound on u_f in (12.4) follows immediately from (ii), since

$$\beta\|u_f\|^2 \leq B(u_f, u_f) = f(u_f) \leq \|f\|_*\|u_f\|.$$

□

Remark 12.11. *Application of Corollary 12.10 to* (WE)*, and use of Lemma 12.8, proves existence and uniqueness of a solution, and establishes that*

$$\|u\| \leq \frac{C_p}{a^-}\|r\|_{L^2}.$$

12.3 Lax–Milgram Theorem

We now relax Corollary 12.10 to apply without the condition of symmetry.

Theorem 12.12 (Lax–Milgram)**.** *Suppose that H is a real Hilbert space and $B : H \times H \to \mathbb{R}$ is a bilinear form satisfying:*

(i) bounded*: there exists $\alpha \geq 0$ such that*

$$|B(u,v)| \leq \alpha\|u\|\|v\| \qquad \textit{for all} \qquad u, v \in H;$$

(ii) coercive*: there exists $\beta > 0$ such that*

$$B(u,u) \geq \beta\|u\|^2 \qquad \textit{for all} \qquad u \in H.$$

Then for any $f \in H^*$, there exists a unique $u_f \in H$ solving (A)*. Furthermore*

$$\|u_f\| \leq \beta^{-1}\|f\|_*; \tag{12.5}$$

in particular, u_f depends continuously on f, i.e.,

$$\|u_f - u_g\| \leq \beta^{-1}\|f - g\|_* .$$

Proof Once we have a solution it is clearly unique: if

$$B(u,v) = B(w,v) = f(v)$$

for every $v \in H$ then $B(u-w,v) = 0$ for every $v \in H$ (since B is bilinear) and in particular for $v = u - w$, whence

$$\beta\|u-w\|^2 \le B(u-w,u-w) = 0,$$

i.e., $u = w$. The bound in (12.5) follows as before (set $v = u_f$ in (A)) and the continuity result follows by considering

$$B(u_f,v) - B(u_g,v) = B(u_f - u_g, v) = (f-g)(v)$$

and setting $v = u_f - u_g$. So only existence requires any work.

Fix $u \in H$, and consider the map $v \mapsto B(u,v)$. We claim that this defines a bounded linear functional on H: it is clearly linear, since

$$B(u,\alpha v_1 + \beta v_2) = \alpha B(u,v_1) + \beta B(u,v_2)$$

by the linearity of B, and it is bounded since

$$|B(u,v)| \le \alpha\|u\|\,\|v\|.$$

It follows from the Riesz Representation Theorem that there exists a $w \in H$ such that

$$\langle w,v\rangle = B(u,v) \qquad \text{for all} \qquad v \in H.$$

We *define* $Au = w$ by the identity

$$\langle Au,v\rangle = B(u,v), \qquad \text{for all} \qquad v \in H$$

and now demonstrate that this definition yields a bounded linear operator from H into itself. Indeed, for every $v \in H$,

$$\begin{aligned}\langle A(\alpha u_1 + \beta u_2), v\rangle &= B(\alpha u_1 + \beta u_2, v)\\ &= \alpha B(u_1,v) + \beta B(u_2,v)\\ &= \alpha\langle Au_1,v\rangle + \beta\langle Au_2,v\rangle\\ &= \langle \alpha Au_1 + \beta Au_2, v\rangle.\end{aligned}$$

Since this holds for every $v \in H$, it follows[2] that

$$A(\alpha u_1 + \beta u_2) = \alpha Au_1 + \beta Au_2,$$

[2] If $\langle u,v\rangle = \langle w,v\rangle$ for every $v \in H$, then $\langle u-w,v\rangle = 0$ for every $v \in H$, in particular, for $v = u - w$, whence $u = w$.

i.e., A is linear. To show that A is bounded, note that

$$\|Au\|^2 = \langle Au, Au\rangle = B(u, Au) \le \alpha\|u\|\|Au\|,$$

so that $\|Au\| \le \alpha\|u\|$ and hence A is bounded.

Using the Riesz Representation Theorem in a standard way, we know that there exists a $\varphi \in H$ such that

$$\langle \varphi, v\rangle = f(v) \qquad \text{for all} \qquad v \in H.$$

We can therefore rewrite our equation as

$$\langle Au, v\rangle = \langle \varphi, v\rangle \qquad \text{for all} \qquad v \in H.$$

This implies that u satisfies (A) if and only if $Au = \varphi$; thus we have reformulated the original problem into a linear operator equation. It remains to show that this equation has a solution. We use coercivity to demonstrate that A is invertible. Now, not only is A "bounded above" but coercivity implies that it is also bounded from below:

$$\beta\|u\|^2 \le B(u, u) = \langle Au, u\rangle \le \|Au\|\|u\| \qquad \Rightarrow \qquad \beta\|u\| \le \|Au\|.$$

As a consequence, A is one-to-one (if $Au = Av$, then $u = v$ since $\|u - v\| \le \beta^{-1}\|Au - Av\| = 0$).

To show that A is onto, first note that

$$\mathrm{Ran}(A) = \{Au : u \in H\}$$

is a closed subspace of H. It is clearly a vector subspace, and it is closed since if $v_n \in \mathrm{Ran}(A)$ (so that $v_n = Au_n$, $u_n \in H$) and $v_n \to v$, then

$$\|u_n - u_m\| \le \beta^{-1}\|Au_n - Au_m\| = \beta^{-1}\|v_n - v_m\|,$$

so that $\{u_n\}$ is Cauchy. Since H is complete, $u_n \to u \in H$, and since A is bounded, it is continuous: it follows that $Au = v$, i.e., $v \in \mathrm{Ran}(A)$, so $\mathrm{Ran}(A)$ is closed. Now suppose that $\mathrm{Ran}(A) \ne H$. It follows that there exists a non-zero $w \in H$ with $w \in (\mathrm{Ran}(A))^\perp$. Thus

$$\beta\|w\|^2 \le B(w, w) = (Aw, w) = 0,$$

a contradiction. Since A is one-to-one and onto, it must have an inverse, i.e., we can find a solution of $Au = \varphi$ for any $\varphi \in H$. □

Example 12.13. Let $D \in \mathbb{R}^{n\times n}$ be strictly positive definite: there is $d_- > 0$ such that for all $v \in \mathbb{R}^n$,

$$\langle Dv, v\rangle \ge d_-\|v\|^2$$

and let $A \in \mathbb{R}^{n\times n}$ be positive semi-definite: for all $v \in \mathbb{R}^n$,

$$\langle Av, v\rangle \geq 0.$$

Note in particular that we do not require D and A to be symmetric.

Assume that, for $r \in \mathbb{R}^n$ and $\epsilon > 0$, u solves the equation

$$(A + \epsilon^{-1}D)u = r.$$

We will show that a unique solution exists for all $\epsilon > 0$ and that $u = u(\epsilon) \to 0$ as $\epsilon \to 0$. Define

$$B(u, v) = \langle (A + \epsilon^{-1}D)u, v\rangle.$$

Then

$$|B(u, v)| \leq \|A + \epsilon^{-1}D\|\|u\|\|v\|$$

where $\|\cdot\|$ denotes both the Euclidean norm on $\mathbb{R}^n$ and the induced operator norm. Note that

$$\begin{aligned} B(u, u) &= \langle Au, u\rangle + \epsilon^{-1}\langle Du, u\rangle \\ &\geq \epsilon^{-1}\langle Du, u\rangle \\ &\geq \epsilon^{-1}d_-\|u\|^2 \end{aligned}$$

where d_- is defined above. We want to find $u = u(\epsilon)$ such that

$$B(u, v) = f(v) \quad \forall v \in \mathbb{R}^n,$$

where $f(v) = \langle r, v\rangle$. Note that

$$\|f\|_{(\mathbb{R}^n)^*} = \sup_{\|v\|=1} |\langle r, v\rangle| \leq \|r\|.$$

Thus by Lax–Milgram we have existence and uniqueness and further:

$$\|u(\epsilon)\| \leq \frac{\|r\|}{\epsilon^{-1}d_-} = \frac{\epsilon\|r\|}{d_-} \to 0 \quad \text{as } \epsilon \to 0.$$

◊

12.4 ODE Analogue

Here we consider an ODE analogue to equation (E) which will be useful for us in several exercises at the end of this lecture. We will use it for exposition on the finite element method (FEM) and several theoretical concepts; however, all of the theory and numerical methods extend naturally to a more complex PDE setting.

Consider the boundary value problem

$$\begin{aligned} -\frac{\mathrm{d}}{\mathrm{d}x}\left(\kappa(x)\frac{\mathrm{d}p}{\mathrm{d}x}(x)\right) &= f(x), \quad x \in (0,1) \\ p(0) = p(1) &= 0. \end{aligned} \tag{BVP}$$

This equation has a natural extension to more general domains $D \subseteq \mathbb{R}^d$, $d > 1$ (say, $d = 2$ or $d = 3$), given by

$$\begin{aligned} -\nabla \cdot \big(\kappa(\boldsymbol{x})\nabla p(\boldsymbol{x})\big) &= f(\boldsymbol{x}), & \boldsymbol{x} &\in D \\ p(\boldsymbol{x}) &= 0, & \boldsymbol{x} &\in \partial D, \end{aligned}$$

where now $\kappa\colon D \to \mathbb{R}^{d\times d}$ is matrix-valued. The theory we look at generalizes to this case directly, and it is in this case where weak solutions and the finite element method become key tools.

From an applications point of view, this equation arises, for example, in groundwater flow models, with κ representing the permeability of the subsurface and p the pressure of a fluid in the subsurface. The matrix-valued function κ is considered to account for the permeability of a medium being dependent on direction (similar to the stress tensor in the Navier–Stokes equations). The equation is arrived at by combining Darcy's law, relating velocity to pressure, with mass conservation. The equation also comes up in electrodynamics: if an electric field E is conservative, then the equation follows from the steady-state Maxwell equations, with κ representing the background conductivity and p representing the electric potential corresponding to E.

For exposition, in Exercises 12.3, 12.4, 12.5, 12.6 and 12.7, we work only with the one-dimensional case $D \equiv (0,1)$ given by (BVP). This allows for the general theory and ideas to be introduced, without having to deal with certain technicalities. Moreover, computations in this case are much faster than those in higher spatial dimensions and easier to implement.

Let us make the following assumptions about (BVP):

Assumptions 12.14. *The diffusion coefficient $\kappa\colon D \to \mathbb{R}$ and source term $f\colon D \to \mathbb{R}$ are such that:*

(i) there exist $\kappa_0, \kappa_1 > 0$ with $\kappa_0 \le \kappa(x) \le \kappa_1$ for all $x \in D$;
(ii) $f \in L^2(0,1) =: L^2(D;\mathbb{R})$.

Exercise 12.3 revisits the C_c^∞ and H_0^1 spaces, weak solutions and connections between them. Exercise 12.4 introduces both theoretical and numerical aspects of FEM. Exercise 12.5 introduces a quadrature solution to equation (BVP). In exercise 12.6, we study convergence

rates of FEM implementations using the quadrature solution from 12.5. Finally, exercise 12.7 explores theoretical properties of equation (BVP) for constant κ and its FEM discretization.

Exercises

12.1(a) Define the operator A acting on functions from $[0,1]$ into $\mathbb{R}$ by

$$(Au)(x) = \int_0^x u(y)dy.$$

Show that if $u \in L^2([0,1];\mathbb{R})$, then $Au \in L^2([0,1];\mathbb{R})$ and that A is a bounded linear operator from $L^2([0,1];\mathbb{R})$ into itself.

(b) Define the operator B acting on functions from $[0,1]$ into $\mathbb{R}$ by

$$(Bu)(x) = \int_x^1 u(y)dy.$$

Show that B is the adjoint of A with respect to the $L^2([0,1];\mathbb{R})$ inner product.

(c) Imagine we are given a function r as data, possibly corrupted by noise, and we want to differentiate it; however, because of noise in the data it may be unwise to differentiate r directly. Instead, we try to find a function u which minimizes

$$I(u) = \frac{1}{2}\|Au - r\|_{L^2}^2 + \frac{\lambda}{2}\|u\|_{L^2}^2.$$

Why, when λ is small and positive, is this a good approach to the problem we wish to solve?

(d) Use the Riesz representation theorem to show that the functional $I\colon L^2([0,1];\mathbb{R}) \to \mathbb{R}^+$ has a unique minimizer.

12.2 Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and let $U \subset H$ be a subspace of H. Let $f : U \to \mathbb{K}$ be a continuous linear functional with $|f(u)| \leq M\|u\|$ for all $u \in U$.

(a) Denote by $\overline{U}$ the closure of U, that is, the union of U and its limit points. Verify that $(\overline{U}, \langle\cdot,\cdot\rangle)$ is a Hilbert space.

(b) Let $\overline{f} : \overline{U} \to \mathbb{K}$ denote the continuous linear extension of f to $\overline{U}$. Show that there exists $v \in \overline{U}$ such that $\|v\| \leq M$ and $\overline{f}(u) = \langle v, u\rangle$ for all $u \in \overline{U}$.

(c) Deduce that there exists $F \in H^*$ such that $F(u) = f(u)$ and $|F(u)| \leq M\|u\|$ for all $u \in U$.

12.3 *This is the first of five exercises on finite element methods. See Section 12.4 for reference and definitions.*

(a) Recall the space $C_c^\infty(0,1)$ of smooth compactly supported functions defined on the open interval $(0,1)$ (Lecture 6). This formulation will be useful:

$$C_c^\infty(0,1) = \left\{ \varphi \in C^\infty(0,1) \;\middle|\; \begin{array}{l} \text{there exists a closed bounded set } K \subset (0,1) \\ \text{with } \varphi(x) = 0 \text{ for all } x \notin K \end{array} \right\}.$$

(i) Let $\varphi \in C_c^\infty(0,1)$. Show that

$$\lim_{x \downarrow 0} \varphi(x) = \lim_{x \uparrow 1} \varphi(x) = 0.$$

(ii) Let p be a solution to (BVP). Show that

$$\int_0^1 \kappa(x) \frac{\mathrm{d}p}{\mathrm{d}x}(x) \frac{\mathrm{d}\varphi}{\mathrm{d}x}(x)\,\mathrm{d}x = \int_0^1 f(x)\varphi(x)\,\mathrm{d}x \quad \forall \varphi \in C_c^\infty(0,1).$$

Do not solve (BVP) explicitly.

(b) Recall the space $H_0^1(0,1)$ which is defined as the completion of $C_c^\infty(0,1)$ w.r.t. the H^1 norm (Lecture 8). Recall that when equipped with H_0^1 norm and inner product:

$$\|\varphi\|_{H_0^1}^2 = \int_0^1 \left| \frac{\mathrm{d}\varphi}{\mathrm{d}x}(x) \right|^2 \mathrm{d}x, \qquad \langle \varphi, \psi \rangle_{H_0^1} = \int_0^1 \frac{\mathrm{d}\varphi}{\mathrm{d}x}(x) \frac{\mathrm{d}\psi}{\mathrm{d}x}(x)\,\mathrm{d}x,$$

H_0^1 is a Hilbert space. In the following, you may use the fact that any $v \in H_0^1(0,1)$ is once differentiable at almost every point in $(0,1)$; in particular, the derivative of v is well-defined whenever it appears under the integral sign.

(i) Prove the Poincaré inequality: for any $v \in H_0^1(0,1)$,

$$\int_0^1 |v(x)|^2\,\mathrm{d}x \le \int_0^1 \left| \frac{\mathrm{d}v}{\mathrm{d}x}(x) \right|^2 \mathrm{d}x.$$

Deduce that $H_0^1(0,1)$ is continuously embedded in $L^2(0,1)$.

(ii) Let Assumptions 12.14 hold, and let p be a solution to (BVP). Using (a)(ii) of the current exercise and the Poincaré inequality, show that

$$\int_0^1 \kappa(x) \frac{\mathrm{d}p}{\mathrm{d}x}(x) \frac{\mathrm{d}v}{\mathrm{d}x}(x)\,\mathrm{d}x = \int_0^1 f(x)v(x)\,\mathrm{d}x \quad \forall v \in H_0^1(0,1). \tag{WBVP}$$

In Exercises 12.4,12.5,12.6,12.7 we will call any function $p \in H_0^1(0,1)$ that satisfies (WBVP) a *weak solution* to (BVP), and any function that solves (BVP) in the classical sense a *strong solution* to (BVP). Note that if it exists, a strong solution is a weak solution. Denote $V \coloneqq H_0^1(0,1)$. Define $B\colon V \times V \to \mathbb{R}$, $g\colon V \to \mathbb{R}$ by

$$B(u,v) = \int_0^1 \kappa(x)\frac{\mathrm{d}u}{\mathrm{d}x}(x)\frac{\mathrm{d}v}{\mathrm{d}x}(x)\,\mathrm{d}x, \qquad g(v) = \int_0^1 f(x)v(x)\,\mathrm{d}x. \tag{12.6}$$

Then we may rewrite (WBVP) in the form

$$B(p,v) = g(v) \qquad \forall v \in V.$$

Using the Lax–Milgram theorem, analogously to Lecture 12, one can show the existence of a unique weak solution $p \in V$ to this equation (Assumptions 12.14 are necessary); however, doing so is not part of this exercise.

12.4 *This is the second of five exercises on finite element methods. See Section 12.4 and Exercise 12.3 for reference and definitions. You may use any programming language of your choice, ignoring explicit MATLAB instructions.*

If we wish to solve (WBVP) numerically, we will first need to make a finite-dimensional approximation to the system, which will then provide a finite-dimensional approximation to the solution. A class of approximations we consider here are known as *finite element approximations*.

Let $V^h \subseteq V$ be a finite-dimensional subspace of V, with basis $\{\varphi_j^h\}_{j=1}^{N_h}$. The scalar parameter $h > 0$ will be related to the dimension of V^h, with $N_h \uparrow \infty$ as $h \downarrow 0$. It is this space V^h in which we will look for an approximate solution.

(a)(i) Equip V^h with the H_0^1 inner product. Show that V^h is a Hilbert space.

(ii) Let $B\colon V^h \times V^h \to \mathbb{R}$, $g\colon V^h \to \mathbb{R}$ denote the restrictions of B, g to V^h, where B, g are defined by (12.6). Using the Lax–Milgram theorem, it can be shown that there exists a unique solution $p^h \in V^h$ to the problem

$$B(p^h, v^h) = g(v^h) \quad \text{for all } v^h \in V^h. \tag{WBVP-h}$$

Let $p^h \in V^h$ denote this solution. Since $\varphi_1^h, \ldots, \varphi_{N_h}^h$ form a

basis for V^h, there exist scalars $P_1^h, \dots, P_{N_h}^h \in \mathbb{R}$ such that

$$p^h = \sum_{j=1}^{N_h} P_j^h \varphi_j^h.$$

Find expressions for the entries of a matrix $A^h \in \mathbb{R}^{N_h \times N_h}$ and a vector $F^h \in \mathbb{R}^{N_h}$ such that the coefficients $P^h = (P_1^h, \dots, P_{N_h}^h)^\top \in \mathbb{R}^{N_h}$ solve

$$A^h P^h = F^h. \tag{FEM}$$

(iii) Define the inner product $\langle \cdot, \cdot \rangle_B$ on V by $\langle u, v \rangle_B = B(u, v)$. Show that the induced norm $\| \cdot \|_B$ is equivalent to $\| \cdot \|_{H_0^1}$.

(iv) Let $p \in V$, $p^h \in V^h$ be the unique solutions to (WBVP) and (WBVP-h), respectively. Establish the *Galerkin orthogonality*:

$$\left\langle p - p^h, v^h \right\rangle_B = 0 \quad \forall v^h \in V^h.$$

Using this, show that

$$\|p - p^h\|_B \leq \|p - v^h\|_B \quad \forall v^h \in V^h$$

and hence

$$\|p - p^h\|_{H_0^1} \leq \sqrt{\frac{\kappa_1}{\kappa_0}} \inf_{v^h \in V^h} \|p - v^h\|_{H_0^1}.$$

The first inequality above is the *Galerkin optimality property*, namely that p^h is optimal over all possible approximations in V^h with respect to the norm induced by B.

(b) We now look at a particular choice of approximation space V^h. Let $h \in (0, 1/2)$ and set $N_h = \lfloor 1/h \rfloor - 1 \in \mathbb{N}$. Define the mesh

$$\boldsymbol{x}^h = \{0, h, 2h, 3h, \dots, N_h h, (N_h + 1)h\} \subset [0, 1].$$

We will denote $x_j^h = jh$ for each $j = 0, \dots, N_h + 1$. The points x_j^h will be referred to as *nodes* and the intervals (x_j^h, x_{j+1}^h) as *elements*.

Define the set of functions $\{\varphi_j^h\} \subseteq H_0^1(0,1)$ by

$$\varphi_j^h(s) = \begin{cases} \dfrac{s - x_{j-1}^h}{x_j^h - x_{j-1}^h} & s \in (x_{j-1}^h, x_j^h], \\ \dfrac{x_{j+1}^h - s}{x_{j+1}^h - x_j^h} & s \in (x_j^h, x_{j+1}^h], \\ 0 & s \notin (x_{j-1}^h, x_{j+1}^h], \end{cases}$$

for each $j = 1, \ldots, N_h$. Define the finite element space

$$V^h = \operatorname{span}\{\varphi_1^h, \ldots, \varphi_{N_h}^h\} \subset H_0^1(0,1).$$

(i) Draw the graphs of each φ_j^h for $h = 1/5$ (you may combine them in one plot, but use color or some other means to distinguish them).

(ii) Show that $\varphi_1^h, \ldots, \varphi_{N_h}^h$ form a basis for V^h. Are they an orthonormal basis?

(iii) Given a function $b \in V$, describe (in words, what it looks like) the function $b^h \in V^h$ given by

$$b^h(x) = \sum_{j=1}^{N_h} b(x_j^h)\varphi_j^h(x).$$

(iv) Observe that φ_j^h and φ_k^h have disjoint supports if $|j - k| > 1$. Given that this is the case, what structure should the matrix A^h in (FEM) have?

(v) Implement the system (FEM) with the terms h, κ, and f provided as parameters. That is, given $h > 0$ and function handles for κ and f, construct A^h and F^h and return the vector of coefficients P^h. It will be beneficial to implement P^h as a sparse matrix (for example, in MATLAB, you may use the `sparse` function).

For integration over elements, you may use the `integral` function in MATLAB (or an equivalent in your language of choice), supplying relative and absolute tolerances of 10^{-14}. We choose these tolerances close to machine precision so that the errors in the integration will be dominated by the errors that arise from the finite element approximation.

Consider the case

$$\kappa(x) = e^x, \quad f(x) = 4e^x(2x + 1).$$

It can be shown that the exact solution is given by the quadratic

$$p(x) = 1 - (2x - 1)^2.$$

Test your implementation in this case by verifying that

$$P_j^h \approx p(x_j^h), \quad j = 1, \ldots, N_h$$

for $h = 2^{-10}$. Construct p^h on the mesh $\boldsymbol{x}^h$ using the coefficients P^h, and produce a plot on logarithmic axes of the error function $e^h(x) = |p^h(x) - p(x)|$.

Remark 12.15. *The definitions we give of the basis $\{\varphi_j^h\}$ can be simplified for the mesh $\boldsymbol{x}^h$ that we work with. Note, however, that the definition makes sense for more general, non-uniform meshes, and the above still applies directly in those cases. Such meshes may be more appropriate to use if either κ or f have strong local behavior that needs to be captured.*

12.5 *This is the third of five exercises on finite element methods. See Section 12.4 and Exercises 12.3, 12.4 for reference and definitions. You may use any programming language of your choice, ignoring explicit MATLAB instructions.*

Consider equation (BVP). As we are in one dimension, we can solve it directly by hand. This allows us to implement a reference solution against which the finite element approximations from Exercise 12.4 can be compared.

By assuming that $f(x) = F'(x)$ for some $F\colon [0,1] \to \mathbb{R}$ and integrating (BVP) twice, it can be shown that the solution to (BVP) is given by

$$p(x) = -\int_0^x \frac{F(y)}{\kappa(y)}\,\mathrm{d}y + \int_0^1 \frac{F(y)}{\kappa(y)}\,\mathrm{d}y \left(\int_0^1 \frac{1}{\kappa(y)}\,\mathrm{d}y\right)^{-1} \int_0^x \frac{1}{\kappa(y)}\,\mathrm{d}y.$$

Fix $h = h_* \coloneqq 2^{-14}$. Using the `integral` function in MATLAB (or an equivalent in your language of choice), implement the solution p of (BVP) on the mesh $\boldsymbol{x}^h$ using the above expression. When calling `integral`, supply relative and absolute tolerances of 10^{-14}.

Test your implementation using the same κ and f as used when testing the finite element method, noting that $f(x) = F'(x)$, where $F(x) = 4e^x(2x - 1)$. Produce a plot on logarithmic axes of the error function for this approximation.

12.6 *This is the fourth of five exercises on finite element methods. See Section 12.4 and Exercises 12.3, 12.4, 12.5 for reference and definitions. You may use any programming language of your choice, ignoring explicit MATLAB instructions.*

Here we study the rates of convergence of the finite element approximations as $h \to 0$, i.e., as the number of basis elements increases. Throughout this question, we fix

$$\kappa(x) = 1.1 + \sin(25x^2), \quad f(x) = \cos(x).$$

We saw in Exercise 12.4(a)(iv) that the error between the true solution and the finite element approximation satisfies the following bound in the H_0^1-norm:

$$\|p - p^h\|_{H_0^1} \leq \sqrt{\frac{\kappa_1}{\kappa_0}} \inf_{v^h \in V^h} \|p - v^h\|_{H_0^1}.$$

Choosing v^h to be the piecewise linear interpolant of the true solution p at the mesh points, which belongs to V^h, it is known that there exists a constant $C(p) > 0$ such that

$$\|p - v^h\|_{H_0^1} \leq C(p)h.$$

We can hence deduce that the convergence of p^h to p as $h \to 0$ is at least linearly fast in the H_0^1-norm. We will see how this compares with numerical experiments and also look at convergence in the L^2-norm.

(a) Denote by p_* the solution to (BVP) using the implementation from Exercise 12.5. Define $h_k := 2^{-k}$. Compute the coefficients P^{h_k} of the finite element approximation p^{h_k} for $k = 4, \ldots, 12$. Construct each p^{h_k} on the mesh $\boldsymbol{x}^{h_*}$ using these coefficients, then calculate and store the H_0^1 and L^2 errors,

$$e_k = \|p_* - p^{h_k}\|_{H_0^1}, \quad e'_k = \|p_* - p^{h_k}\|_{L^2},$$

using the `integral`, `gradient`, and `interp1` functions in MATLAB (or equivalents in other languages). It may also be beneficial to run `format long` to increase the number of decimal places in the output. Present these errors in a table: you will have three columns: k, e_k, and e'_k, and 9 rows for $k = 4, \ldots, 12$.

(b) Given a norm $\|\cdot\|$ and $h, h' > 0$, we define the *experimental order of convergence* (EOC) with respect to $\|\cdot\|$ by

$$\mathrm{EOC}(h, h') = \frac{\log\left(\|p - p^h\| / \|p - p^{h'}\|\right)}{\log(h/h')}.$$

(i) Explain why if the convergence is of the form

$$\|p - p^h\| \approx Ch^\alpha,$$

then $\mathrm{EOC}(h, h')$ provides an estimate for α.

(ii) For each of the H_0^1 and L^2-norms, compute $\mathrm{EOC}(h_k, h_{k+1})$ for $k = 4, \ldots, 11$ using the stored values of e_k, e_k', and present these values in a table (3 columns and 8 rows).
How do the experimental rates of convergence in the H_0^1-norm compare with the theoretical bound above? What do you think the theoretical rate of convergence in the L^2-norm is?

Remark 12.16. *Faster rates of convergence can be attained by choosing a different family of finite element basis functions $\{\varphi_j^h\}$. Rather than being piecewise linear, they can be chosen to be piecewise quadratic, piecewise cubic, and so on. These basis functions have a larger support than the piecewise linear basis functions, and so the matrix P^h becomes less sparse. Thus there is an increase in, for example, memory requirements in exchange for the higher rate of convergence.*

12.7 *This is the fifth of five exercises on finite element methods. See Section 12.4 and Exercises 12.3, 12.4, for reference and definitions. You may use any programming language of your choice, ignoring explicit MATLAB instructions.*

In this exercise, we focus on the case where the diffusion coefficient κ is constant. The equation (BVP) is then known as the Poisson equation.

(a) Fix $\kappa = 1$. Let L denote the differential operator associated with the left-hand side of (BVP), i.e., the negative second derivative operator equipped with zero boundary conditions on $(0, 1)$.

(i) Find the eigenfunctions $\{\psi_j\}$ and corresponding eigenvalues $\{\lambda_j\}$ associated with L:

$$-\frac{\mathrm{d}^2\psi_j}{\mathrm{d}x^2}(x) = \lambda_j\psi_j(x), \quad x \in (0, 1),$$
$$\psi_j(0) = \psi_j(1) = 0.$$

(ii) Show that the bound in the Poincaré inequality from Exercise 12.3(b)(i) may be improved to

$$\int_0^1 |v(x)|^2 \, \mathrm{d}x \leq \frac{1}{\pi^2} \int_0^1 \left| \frac{\mathrm{d}v}{\mathrm{d}x}(x) \right|^2 \, \mathrm{d}x.$$

You may use without proof that the eigenfunctions $\{\psi_j\}$ form an orthonormal basis for $L^2(0,1)$.

(b) Fix $\kappa = 1$ and $h > 0$. Define $L^h \in \mathbb{R}^{N_h \times N_h}$ by $L^h = A^h/h$, where A^h denotes the finite element matrix in (FEM) associated with the discretization of (BVP), using the approximation from Exercise 12.4(b). L^h is an approximation to L, often referred to as the *finite difference Laplacian*. The rescaling of A^h by $1/h$ is natural after approximating the entries of the vector F^h with the trapezoidal rule.

(i) Show that the eigenvectors $\{W^{(j)}\}$ and corresponding eigenvalues $\{\mu_j\}$ of L^h satisfy the system

$$\begin{aligned} -W_2^{(j)} + 2W_1^{(j)} &= \mu_j h^2 W_1^{(j)}, \\ -W_{k+1}^{(j)} + 2W_k^{(j)} - W_{k-1}^{(j)} &= \mu_j h^2 W_k^{(j)}, \qquad k = 2, \ldots, N_h - 1, \\ 2W_{N_h}^{(j)} - W_{N_h-1}^{(j)} &= \mu_j h^2 W_{N_h}^{(j)}. \end{aligned}$$

(ii) It can be shown (after some work with trigonometric identities) that, up to normalization constant,

$$W_k^{(j)} = \sin(j\pi k h), \quad k = 1, \ldots, N_h,$$

and the corresponding eigenvalues $\{\mu_j\}$ are given by

$$\mu_j = \frac{4\sin(j\pi h/2)^2}{h^2}.$$

Deduce that the spectrum of L^h may be bounded below by a positive constant independently of h.

(iii) Show that for all $U \in \mathbb{R}^{N_h}$, a discrete analog of the Poincaré inequality holds:

$$\|U\|_2^2 \leq \frac{h^2}{4\sin(\pi h/2)^2} \langle U, L^h U \rangle_2.$$

13
Spectral Theorem

In this lecture we prove the spectral theorem, diagonalization of symmetric matrices, in finite dimensions. We then move on to state and prove the analogous result for compact operators in an infinite-dimensional space. Whilst the spectral theorem may be viewed as diagonalization of a symmetric operator in the orthonormal basis of eigenvectors, it also leads to a representation of the operator as a sum of rank one operators. This representation leads to low rank approximation theorems for the operator.

13.1 Preliminaries

Let $\left(H, \langle \cdot, \cdot \rangle, \|\cdot\|\right)$ be a Hilbert space over $\mathbb{K} = \mathbb{R}$. Although we work in real-valued Hilbert spaces, we will, when discussing the eigenvalue problem, have recourse to the *complexification* of the real-valued Hilbert space. This just allows extension of operators defined on real-valued H to apply to complex-valued vector space elements; this is analogous to applying a real matrix to a complex vector. We omit the full details of the complexification process since they are somewhat lengthy, yet quite intuitive.

Definition 13.1. *Symmetric operator $T \in \mathcal{L}(H, H)$ is* positive *if $\langle h, Th \rangle \geq 0$ for all $h \in H$.*

Definition 13.2. *Let $T \in \mathcal{L}(H, H)$, where here H has been complexified. We define the* eigenvalue/eigenvector problem *as follows: find $(\lambda, u) \in \mathbb{C} \times H$ such that*

$$\begin{aligned} Tu &= \lambda u, \\ \|u\|^2 &= 1. \end{aligned} \tag{EVP-OP}$$

We call $\lambda \in \mathbb{C}$ an eigenvalue *of T and $u \in H$, $u \neq 0$, an* eigenvector. *Together they are referred to as an* eigenpair. *The set of all eigenvalues of T is denoted by $\sigma(T)$.*

Remark 13.3. *We will not refer to the complexification of H again; we will simply complexify (implicitly) when needed.*

Proposition 13.4. *Symmetric operator $T \in \mathcal{L}(H, H)$ has only real eigenvalues and eigenvectors; furthermore, eigenvectors corresponding to distinct eigenvalues are orthogonal. If, in addition, the operator is positive, then the eigenvalues are non-negative.*

Proof Let $(\lambda_i, u^{(i)})$ be eigenvalue/eigenvector pairs of T corresponding to distinct eigenvalues indexed by i. Note that for any single pair (we may then drop the index),

$$\langle u, Tu\rangle = \langle u, \lambda u\rangle = \lambda\|u\|^2,$$
$$\langle Tu, u\rangle = \langle \lambda u, u\rangle = \overline{\lambda}\|u\|^2.$$

The left hand side of both these identities is equal because T is symmetric. It follows that $\lambda = \overline{\lambda}$. It also follows that if T is positive, then $\lambda \geq 0$. Also, for $i \neq j$,

$$\lambda_i\langle u^{(i)}, u^{(j)}\rangle = \langle Tu^{(i)}, u^{(j)}\rangle = \langle u^{(i)}, Tu^{(j)}\rangle = \lambda_j\langle u^{(i)}, u^{(j)}\rangle.$$

Since $\lambda_i \neq \lambda_j$ the orthogonality result follows. □

Example 13.5. While the definition of an eigenvalue problem is familiar from linear algebra, the properties of the set of eigenvalues can be startlingly different in the realm of infinite dimensions. Here we provide two examples of operators, one of which has no eigenvalues at all, and the other with a continuum of eigenvalues.

Consider the right-shift operator from Example 7.8:

$$S_r u = (0, u_1, u_2, u_3, \ldots)$$

for any $u = (u_1, u_2, u_3, \ldots)$. Then $S_r : \ell^2(\mathbb{N}; \mathbb{R}) \to \ell^2(\mathbb{N}; \mathbb{R})$. This operator has no eigenvalues, since $S_r u = \lambda u$ implies that

$$(0, u_1, u_2, \ldots) = \lambda(u_1, u_2, u_3, \ldots),$$

and so

$$\lambda u_1 = 0, \quad \lambda u_2 = u_1, \quad \lambda u_3 = u_2, \quad \ldots$$

If $\lambda \neq 0$, then this implies that $u_1 = 0$, and then $u_2 = u_3 = u_4 = \ldots = 0$, and so λ is not an eigenvalue since an eigenvector must have norm 1.

If $\lambda = 0$, then we also obtain $u = 0$, and so there are no eigenvalues, i.e., $\sigma(S_r) = \emptyset$. This cannot happen in finite dimensions where every matrix has at least one eigenvalue.

On the other hand, consider the left-shift operator $S_l \colon \ell^2(\mathbb{N};\mathbb{R}) \to \ell^2(\mathbb{N};\mathbb{R})$, defined by

$$S_l u = (u_2, u_3, u_4, \ldots)$$

for any $u = (u_1, u_2, u_3, \ldots)$. By definition, $\lambda \in \mathbb{C}$ is an eigenvalue if $S_l u = \lambda u$, i.e., if

$$(u_2, u_3, u_4, \ldots) = \lambda(u_1, u_2, u_3, \ldots),$$

i.e., if

$$u_2 = \lambda u_1, \quad u_3 = \lambda u_2, \quad u_4 = \lambda u_3, \quad \ldots$$

and so on. Given $\lambda \neq 0$, this gives a candidate eigenvector

$$u = (1, \lambda, \lambda^2, \lambda^3, \ldots),$$

which is an element of $\ell^2(\mathbb{N};\mathbb{R})$ provided that

$$\sum_{n=0}^{\infty} |\lambda|^{2n} = \frac{1}{1 - |\lambda|^2} < \infty,$$

which, in turn, is the case for any λ with $|\lambda| < 1$. It follows that

$$\left\{\lambda \in \mathbb{C} : \ |\lambda| < 1\right\} \subseteq \sigma(S_l).$$

Thus the eigenvalues form an uncountable set. This cannot happen in finite dimensions where the set of eigenvalues is finite and defined by the roots of a polynomial.

◊

13.2 Finite Dimensional Spectral Theorem

We now work in $\mathbb{R}^n$: let $A \in \mathbb{R}^{n\times n}$ and $\|\cdot\|$, $\langle \cdot\,, \cdot \rangle$ be Euclidean norm and inner product, respectively. We define the *eigenvalue/eigenvector problem* as follows: find $(\lambda, \varphi) \in \mathbb{C} \times \mathbb{C}^n$ such that

$$\begin{aligned} A\varphi &= \lambda\varphi, \\ \|\varphi\|^2 &= 1. \end{aligned} \tag{EVP}$$

We will index the set of solutions by j: $(\lambda_j, \varphi^{(j)})$.

Proposition 13.6. *If $A \in \mathbb{R}^{n\times n}$ is symmetric ($A^\top = A$), then A has at least one real eigenvalue/eigenvector pair.*

Proof Let $\Psi(x) = \frac{1}{2}\langle x, Ax\rangle$ and note that then, using symmetry of A, $\nabla\Psi(x) = Ax$. Let $S = \{x \in \mathbb{R}^n : \|x\|^2 = 1\}$. S is compact, and so Ψ, as a continuous function on S, attains its maximum in S. To find it, define

$$J(x, \lambda) \coloneqq \Psi(x) - \frac{1}{2}\lambda(\|x\|^2 - 1).$$

Then, the maximum can be found as the solution of

$$\begin{aligned} \nabla\Psi(x) - \lambda x &= 0, \\ \frac{1}{2}(\|x\|^2 - 1) &= 0, \end{aligned}$$

or, equivalently,

$$\begin{aligned} Ax &= \lambda x, \\ \|x\|^2 &= 1. \end{aligned}$$

Hence the solution (λ, x) is a *real* eigenvalue/eigenvector pair.

□

Definition 13.7. *A matrix $Q \in \mathbb{C}^{m\times n}$ with $m \geq n$ is* unitary *if $Q^*Q = I$. If Q is real, then the condition is that $Q^\top Q = I$, and Q is said to be* orthogonal.

Theorem 13.8 (Spectral Theorem in $\mathbb{R}^n$)**.** *The following three statements are equivalent:*

(i) $A \in \mathbb{R}^{n\times n}$ is symmetric;
(ii) there exists an orthonormal basis for $\mathbb{R}^n$ comprising eigenvectors of A;
(iii) there exists P orthogonal such that $P^\top AP = D$, where D is a real diagonal matrix.

Proof
(iii) $\Rightarrow$ (i) By assumption, there exists orthogonal P such that $A = PDP^\top$. Then we may write:

$$A^\top = \left(PDP^\top\right)^\top = \left(P^\top\right)^\top D^\top P^\top = PDP^\top = A.$$

(ii) $\Rightarrow$ (iii) Again, by assumption, there exists an orthonormal eigenbasis:

$$\begin{aligned} Av^{(j)} &= \lambda_j v^{(j)}, \quad j = 1, \ldots, n \\ \left\langle v^{(j)}, v^{(k)} \right\rangle &= \delta_{jk}. \end{aligned} \tag{13.1}$$

Let $P = \left(v^{(1)}, \ldots, v^{(n)}\right)$, then $P^\top P = I$. Also,

$$AP = \left(Av^{(1)}, \ldots, Av^{(n)}\right) = \left(\lambda_1 v^{(1)}, \ldots, \lambda_n v^{(n)}\right) = PD,$$
$$D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}.$$

We conclude that $P^\top AP = D$.

(iii) $\Rightarrow$ (ii) If $AP = PD$ and $P^\top P = I$, then simply define $v^{(j)}$ as columns of P, and λ_j as diagonal entries of matrix D. In other words,

$$P = \left(v^{(1)}, \ldots, v^{(n)}\right),$$
$$D = \operatorname{diag}\{\lambda_1, \ldots, \lambda_n\}.$$

Then we will trivially have (13.1).

(i) $\Rightarrow$ (iii) Proceed by induction. Clearly true for $n = 1$. We assume (i) $\Rightarrow$ (iii) holds for $(n-1) \times (n-1)$ matrices. Now consider a symmetric matrix $A \in \mathbb{R}^{n \times n}$. By Proposition 13.6, there exists $v^{(1)}$, $\|v^{(1)}\|^2 = 1$ and $\lambda_1 \in \mathbb{R}$ such that:

$$Av^{(1)} = \lambda_1 v^{(1)},$$
$$\|v^{(1)}\|^2 = 1.$$

Making use of the Gram–Schmidt orthonormalization process, we may arrive at an orthonormal basis $\left\{v^{(j)}\right\}_{j=1}^n$, with $v^{(1)}$ defined above.

Let $P_1 = \left(v^{(1)}, \ldots, v^{(n)}\right)$, then

$$P_1^\top A P_1 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & B \end{pmatrix}, \quad B \in \mathbb{R}^{(n-1)\times(n-1)}.$$

The first column is $(\lambda_1, 0, \ldots, 0)^\top$ because

$$AP_1 = \left(\lambda_1 v^{(1)}, Av^{(2)}, \ldots, Av^{(n)}\right), \tag{13.2}$$
$$\left\langle v^{(1)}, v^{(j)} \right\rangle = \delta_{1,j}, \quad j = 1, \ldots, n, \tag{13.3}$$

and the remaining structure follows from the symmetry assumption. In particular, B should also be symmetric.

By the inductive hypothesis, there exist an orthogonal matrix P_2 and a real diagonal matrix D_2 such that $B = P_2 D_2 P_2^\top$. Therefore,

$$P_1^\top A P_1 = \begin{pmatrix} \lambda_1 & 0 \\ 0 & P_2 D_2 P_2^\top \end{pmatrix}.$$

Setting

$$P = P_1 \begin{pmatrix} 1 & 0 \\ 0 & P_2 \end{pmatrix}, \quad D = \begin{pmatrix} \lambda_1 & 0 \\ 0 & D_2 \end{pmatrix},$$

we finally arrive at the following expression, concluding the proof:

$$P^\top A P = \begin{pmatrix} 1 & 0 \\ 0 & P_2^\top \end{pmatrix} P_1^\top A P_1 \begin{pmatrix} 1 & 0 \\ 0 & P_2 \end{pmatrix} = \\ \begin{pmatrix} 1 & 0 \\ 0 & P_2^\top \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & B \end{pmatrix} \begin{pmatrix} 1 & 0 \\ 0 & P_2 \end{pmatrix} = \begin{pmatrix} \lambda_1 & 0 \\ 0 & D_2 \end{pmatrix}$$

□

Theorem 13.9. *If $A \in \mathbb{R}^{n \times n}$ is symmetric, then we may write*

$$A = \sum_{j=1}^{n} \lambda_j v^{(j)} \otimes v^{(j)},$$

where $\left(\lambda_j, v^{(j)}\right)$ are real orthonormal eigenvalue/eigenvector pairs.

Proof By the spectral theorem, we have a decomposition $A = PDP^\top$ where P has eigenvectors $\{v^{(j)}\}_{j=1}^n$ as columns, and D is a diagonal matrix with corresponding eigenvalues $\{\lambda_j\}_{j=1}^n$ on the diagonal. Thus, for any $u \in \mathbb{R}^n$,

$$Au = PDP^\top u = PD \begin{pmatrix} \left\langle v^{(1)}, u \right\rangle \\ \vdots \\ \left\langle v^{(n)}, u \right\rangle \end{pmatrix} = P \begin{pmatrix} \lambda_1 \left\langle v^{(1)}, u \right\rangle \\ \vdots \\ \lambda_n \left\langle v^{(n)}, u \right\rangle \end{pmatrix} = \\ \sum_{j=1}^{n} \lambda_j \left\langle v^{(j)}, u \right\rangle v^{(j)} = \left(\sum_{j=1}^{n} \lambda_j v^{(j)} \otimes v^{(j)} \right) u.$$

Since u is arbitrary the desired result follows. □

13.3 Spectral Theory for Compact Symmetric Operators

We now return to the general setting where $\left(H, \langle \cdot, \cdot \rangle, \| \cdot \|\right)$ is a Hilbert space.

Definition 13.10. *Let $A \in \mathcal{L}(H, H)$. Then the* kernel *of A, denoted* $\mathrm{Ker}(A)$*, is defined by*

$$\mathrm{Ker}(A) \coloneqq \{h \in H : Ah = 0\}.$$

Lemma 13.11. *Let $A \in \mathcal{L}(H,H)$ be symmetric. Then*

$$\varsigma(A) \coloneqq \sup_{\|f\|=1} |\langle f, Af\rangle| = \|A\|.$$

Proof Clearly

$$\varsigma(A) \leq \sup_{\|f\|=1} \|A\|\|f\|^2 = \|A\|.$$

We now prove the reverse inequality. A straightforward calculation shows that

$$\langle f_1 + f_2, A(f_1+f_2)\rangle - \langle f_1 - f_2, A(f_1-f_2)\rangle = 4\mathrm{Re}\langle f_2, Af_1\rangle.$$

Thus, dividing and multiplying by $\|f_1 \pm f_2\|^2$ and taking supremum,

$$4\mathrm{Re}\langle f_2, Af_1\rangle \leq \varsigma(A)\Big(\|f_1+f_2\|^2 + \|f_1-f_2\|^2\Big) = 2\varsigma(A)\Big(\|f_1\|^2 + \|f_2\|^2\Big).$$

For $f_1 \notin \mathrm{Ker}(A)$ with $\|f_1\| = 1$, we define $f_2 = Af_1/\|Af_1\|$ noting that the above identity gives

$$4\|Af_1\| \leq 4\varsigma(A),$$

an inequality which is also clearly true for $f_1 \in \mathrm{Ker}(A)$. It follows that

$$\|A\| = \sup_{\|f_1\|=1} \|Af_1\| \leq \varsigma(A),$$

completing the proof. □

Theorem 13.12 (Spectral Theorem)**.** *Let $A \in \mathcal{L}(H,H)$ be compact and symmetric. Then:*

- *all eigenvalues and eigenvectors of A are real;*
- *A has at least one eigenvalue;*
- *A has at most countably many non-zero eigenvalues $\{\lambda_n\}_{n\in\mathbb{N}}$;*
- *the sequence of eigenvalues has no non-zero accumulation points;*
- *there exists an orthonormal basis for $\mathrm{Ker}(A)^\perp$ comprising eigenvectors $\{v^{(n)}\}_{n\in\mathbb{N}}$ satisfying $Av^{(n)} = \lambda_n v^{(n)}$.*

Proof The first item is proved in Proposition 13.4. If $A = 0$, then there is nothing to show, and hence we assume $A \neq 0$ without loss of generality. Let $\{f^{(k)}\}_{k\in\mathbb{N}}$ be a sequence attaining the supremum in Lemma 13.11. Then, as $k \to \infty$,

$$|\langle f^{(k)}, Af^{(k)}\rangle| \to \varsigma(A) = \|A\|.$$

Since A is symmetric, the inner product is real and the limit of $\langle f^{(k)}, Af^{(k)}\rangle$ accumulates at $\pm\|A\|$. Along a subsequence, we have

$$\langle f^{(k)}, Af^{(k)}\rangle \to \lambda_1, \quad k \to \infty, \tag{13.4}$$

with $\lambda_1 \in \{\pm\|A\|\}$. Since the $f^{(k)}$ have norm one, Theorem 12.4 asserts the existence of $v^{(1)} \in H$ with $f^{(k)} \rightharpoonup v^{(1)}$ as $k \to \infty$, by moving to a subsequence. Lemma 10.13 then implies that $Af^{(k)} \to Av^{(1)}$ as $k \to \infty$ since A is compact.

Since A is non-zero, λ_1 is non-zero. Using the weak convergence of $f^{(k)}$ to $v^{(1)}$ and the strong convergence of $A(f^{(k)} - v^{(1)})$ to zero, the display (13.4) shows that

$$\langle v^{(1)}, Av^{(1)}\rangle = \lambda_1. \tag{13.5}$$

Hence $v^{(1)} \neq 0$. Also,

$$\begin{aligned}
\|Af^{(k)} - \lambda_1 f^{(k)}\|^2 &= \|Af^{(k)}\|^2 - 2\lambda_1\langle f^{(k)}, Af^{(k)}\rangle + \lambda_1^2\|f^{(k)}\|^2 \\
&\leq \|A\|^2 - 2\lambda_1\langle f^{(k)}, Af^{(k)}\rangle + \lambda_1^2 \\
&= 2\lambda_1\big(\lambda_1 - \langle f^{(k)}, Af^{(k)}\rangle\big) \\
&\to 0 \quad \text{as} \quad k \to \infty.
\end{aligned}$$

It is a consequence that

$$\lambda_1 f^{(k)} = \big(\lambda_1 f^{(k)} - Af^{(k)}\big) + \big(Af^{(k)} - Av^{(1)}\big) + Av^{(1)} \to Av^{(1)} \quad \text{as} \quad k \to \infty$$

since the terms in parentheses both tend to zero. Thus, for all $h \in H$,

$$\langle \lambda_1 f^{(k)}, h\rangle \to \langle Av^{(1)}, h\rangle.$$

By the weak convergence of $f^{(k)}$ to $v^{(1)}$, we deduce that

$$\langle \lambda_1 f^{(k)}, h\rangle \to \langle \lambda_1 v^{(1)}, h\rangle.$$

Combining shows that

$$\langle Av^{(1)} - \lambda_1 v^{(1)}, h\rangle = 0,$$

and, since $h \in H$ is arbitrary, that $Av^{(1)} = \lambda_1 v^{(1)}$. This, together with (13.5), implies that $\|v^{(1)}\| = 1$, and so we have an eigenvalue-eigenvector pair $(\lambda_1, v^{(1)})$ for A.

We now define $H_2 = \operatorname{span}\{v^{(1)}\}^\perp \subset H$. We notice that, for $f \in H_2$,

$$\langle Af, v^{(1)}\rangle = \langle f, Av^{(1)}\rangle = \lambda_1\langle f, v^{(1)}\rangle = 0$$

so that $A\colon H_2 \to H_2$. Hence define $A_2 := A|_{H_2} \in \mathcal{L}(H_2, H_2)$. Since $H_2 \subset H$, it follows that $\|A_2\| \leq \|A\| = |\lambda_1|$. Unless $A_2 = 0$, in which

case the recursive procedure we now define terminates, we may proceed as above to deduce the existence of $\lambda_2 \in \mathbb{R}$, $v^{(2)} \in H_2 \subset H$ such that $Av^{(2)} = \lambda_2 v^{(2)}$ so that $v^{(2)}$ is a unit norm eigenvector corresponding to eigenvalue λ_2. By construction, $v^{(2)} \perp v^{(1)}$ and $0 < |\lambda_2| \leq |\lambda_1|$.

Proceeding by induction we find

$$\begin{aligned} H &= H_1 \supset H_2 \supset H_3 \supset \cdots, \\ H_n &= \operatorname{span}\{v^{(1)}, \cdots, v^{(n-1)}\}^{\perp}, \\ A_n &= A|_{H_n} \in \mathcal{L}(H_n, H_n), \\ Av^{(n)} &= \lambda_n v^{(n)}, \quad \{v^{(1)}, \cdots, v^{(n)}\} \text{ orthonormal}, \\ |\lambda_n| &= \|A_n\|, \quad |\lambda_1| \geq |\lambda_2| \geq \cdots. \end{aligned}$$

Either the procedure terminates at some $m \in \mathbb{N}$ for which $A_m = 0$, or continues for all $n \in \mathbb{N}$. In the former case, A has finite rank m and has finitely many non-zero eigenvalues. In the latter case, we now assume for contradiction that the λ_n accumulate at a point which is not zero: there exists $\delta > 0$ such that, for all $n \in \mathbb{N}$, $|\lambda_n| \geq \delta > 0$. Because the $\{v^{(n)}\}_{n\in\mathbb{N}}$ are orthonormal, Theorem 12.5 shows that $v^{(n)} \rightharpoonup 0$; hence, because A is compact, Lemma 10.13 shows that $Av^{(n)} \to 0$. But

$$\|Av^{(n)}\| = |\lambda_n|\|v^{(n)}\| \geq \delta > 0,$$

a contradiction.

In the final step of the proof, we show that $\{v^{(n)}\}_{n\in\mathbb{N}}$ forms an orthonormal basis for $\operatorname{Ker}(A)^{\perp}$. When there is $m \in \mathbb{N}$ at which the above recursion terminates, then $\operatorname{Ker}(A)^{\perp}$ has dimension $m-1$ and the algorithm produces an orthonormal set of vectors $\{v^{(n)}\}_{n=1}^{m-1}$ which span $\operatorname{Ker}(A)^{\perp}$ so the desired result is proved. In the case where there is no such m, then $\operatorname{Ker}(A)^{\perp}$ is infinite-dimensional and more care is needed. Considering this case, let f be an arbitrary element in $\operatorname{Ker}(A)^{\perp}$. Define

$$f' = \sum_{n=1}^{\infty} \langle v^{(n)}, f\rangle v^{(n)} \in \operatorname{Ker}(A)^{\perp}. \tag{13.6}$$

Note that, by construction,

$$\langle f', v^{(n)}\rangle = \langle f, v^{(n)}\rangle \quad \forall n \in \mathbb{N},$$

from which it follows that $f - f'$ is orthogonal to $v^{(n)}$ for all $n \in \mathbb{N}$. Thus

$$f - f' \in \bigcap_{n\in\mathbb{N}} H_n.$$

As a consequence, for all $n \in \mathbb{N}$,

$$\begin{aligned}\|A(f-f')\| = \|A_n(f-f')\|_{H_n} &\leq \|A_n\|\|f-f'\|_{H_n} \\ &\leq |\lambda_n|\|f-f'\|_H.\end{aligned}$$

Since $|\lambda_n| \to 0$ as $n \to \infty$, we deduce that $A(f-f') = 0$ so that $f - f' \in \mathrm{Ker}(A)$. But, by design, $f - f' \in \mathrm{Ker}(A)^\perp$ and so $f - f' = 0$. It follows that $f = f'$ and hence that $\{v^{(n)}\}_{n\in\mathbb{N}}$ is indeed an orthonormal basis for $\mathrm{Ker}(A)^\perp$ and that $f = f'$ may be represented in this basis by means of formula (13.6). □

Consider now the setting in which there is no finite terminating m, so that $\mathrm{Ker}(A)^\perp$ is infinite-dimensional. Then any $f \in \mathrm{Ker}(A)^\perp$ can be expanded as

$$f = \sum_{n=1}^{\infty} \langle v^{(n)}, f\rangle v^{(n)}.$$

Since A is continuous, we may write

$$\begin{aligned}Af &= \sum_{n=1}^{\infty} \langle v^{(n)}, f\rangle A v^{(n)}, \\ &= \sum_{n=1}^{\infty} \lambda_n \langle v^{(n)}, f\rangle v^{(n)}.\end{aligned}$$

This holds for any $f \in H$ since the part of f in $\mathrm{Ker}(A)$ disappears under the action of A. We thus have established:

Corollary 13.13. *Let $A \in \mathcal{L}(H,H)$ be compact and symmetric. Then*

$$A = \sum_{n=1}^{\infty} \lambda_n v^{(n)} \otimes v^{(n)}, \tag{13.7}$$

where $\left(\lambda_n, v^{(n)}\right)$ are the real orthonormal eigenvalue/eigenvector pairs from Theorem 13.12.

In the next section, we discuss approximation of this infinite series representation of A.

13.4 Approximation of Compact Symmetric Operators

As in the previous section, we let $\left(H, \langle\cdot,\cdot\rangle, \|\cdot\|\right)$ be a Hilbert space.

Definition 13.14. *Let $T : H \to H$. Then the* range *or* image *of T is the set*

$$T(H) := \{v \in H : \exists u \in H \text{ with } v = Tu\}.$$

This is sometimes denoted $\mathrm{Ran}(T)$ *or* $\mathrm{Im}(T)$.

Remark 13.15. *Recalling Definition 7.13, we observe that* $\operatorname{rank} T = \dim T(H)$, *where* $\dim T(H)$ *denotes the number of linearly independent elements of H which span $T(H)$.*

Consider the compact symmetric operator A from the previous section. If there is finite m at which the recursion terminates, then $\operatorname{rank} A = m$; otherwise, the rank is infinite. In this section we discuss the setting where we approximate an infinite rank operator by one with finite rank. In view of formula (13.7), a natural rank r approximation is

$$A_r := \sum_{n=1}^{r} \lambda_n v^{(n)} \otimes v^{(n)}. \tag{13.8}$$

We recall the following definition from Exercise 10.8:

Definition 13.16. *Operator $T \in \mathcal{L}(H, H)$ is Hilbert-Schmidt if, for some orthonormal basis $\{e^{(n)}\}_{n\in\mathbb{N}}$,*

$$\|T\|_{HS} := \Big(\sum_{n=1}^{\infty} \|Te^{(n)}\|^2\Big)^{\frac{1}{2}} < \infty.$$

Lemma 13.17. *The preceding definition of $\|\cdot\|_{HS}$ is independent of the choice of orthonormal basis $\{e^{(n)}\}_{n\in\mathbb{N}}$ and defines a norm on $\mathcal{L}(H, H)$.*

Example 13.18. Let $A \in \mathcal{L}(H, H)$ be compact and symmetric, in the setting of the previous section in which $\mathrm{Ker}(A)^\perp$ is infinite-dimensional. Then, choosing the eigenbasis $\{e^{(n)} = v^{(n)}\}_{n\in\mathbb{N}}$,

$$\|A\|_{HS} = \Big(\sum_{n=1}^{\infty} \lambda_n^2\Big)^{\frac{1}{2}}.$$

Thus compact symmetric A is Hilbert-Schmidt if and only if the ordered eigenvalue sequence is an element of $\ell^2(\mathbb{N}; \mathbb{R})$. ◊

Theorem 13.19. *Let $A \in \mathcal{L}(H, H)$ be compact, symmetric and Hilbert-Schmidt, and assume that* $\mathrm{Ker}(A)^\perp$ *is infinite-dimensional. Then the finite rank approximation A_r given by* (13.8) *converges to A given by* (13.7) *in the sense that $\|A - A_r\|_{HS} \to 0$ as $r \to \infty$.*

Proof It is straightforward that

$$A - A_r = \sum_{n=r+1}^{\infty} \lambda_n v^{(n)} \otimes v^{(n)}.$$

Furthermore

$$(A - A_r)v^{(\ell)} = \sum_{n=r+1}^{\infty} \lambda_n v^{(n)} \langle v^{(n)}, v^{(\ell)} \rangle = \lambda_\ell v^{(\ell)},$$

for $\ell \geq r+1$ and is zero otherwise. Thus

$$\|A - A_r\|_{HS}^2 = \sum_{\ell=r+1}^{\infty} \lambda_\ell^2.$$

Because A is Hilbert-Schmidt, we have

$$\sum_{\ell=1}^{\infty} \lambda_\ell^2 < \infty$$

from which it follows that

$$\sum_{\ell=r+1}^{\infty} \lambda_\ell^2 \to 0$$

as $r \to \infty$, and hence the result is proved.

□

Example 13.20. Consider $H = \ell^2(\mathbb{N}; \mathbb{R})$, and suppose we are given $k : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$, that is, $\{k_{ij}\}_{(i,j)\in\mathbb{N}\times\mathbb{N}}$, with $k_{ij} = k_{ji}$ and there exists $s > 0$ such that

$$\sum_{(i,j)\in\mathbb{N}\times\mathbb{N}} (i^s k_{ij})^2 = K^+ < \infty.$$

Define operator $K \in \mathcal{L}(H, H)$ by the following relation:

$$u \mapsto Ku = v, \qquad v_i = \sum_{j\in\mathbb{N}} k_{ij} u_j.$$

The operator K is linear and symmetric; we now would like to show that it is compact.

First, recall the definition of the $\mathcal{X}^s$ space:

$$\mathcal{X}^s := \left\{ h \in H : \|h\|_{\mathcal{X}^s} = \left(\sum_{j\in\mathbb{N}} j^{2s} h_j^2 \right)^{1/2} < \infty \right\}$$

In Lecture 10, we proved that $\mathcal{X}^s$ is compactly embedded in ℓ^2, for any $s > 0$. Therefore, K is compact if it maps a bounded set in ℓ^2 to a bounded set in $\mathcal{X}^s$.

We have:

$$\|v\|_{\mathcal{X}^s}^2 = \sum_{i\in\mathbb{N}} i^{2s} v_i^2 = \sum_{i\in\mathbb{N}} i^{2s} \left(\sum_{j\in\mathbb{N}} k_{ij} u_j \right)^2 \leq \sum_{i\in\mathbb{N}} i^{2s} \left(\sum_{j\in\mathbb{N}} k_{ij}^2 \right) \left(\sum_{j\in\mathbb{N}} u_j^2 \right),$$

where we applied Cauchy–Schwarz to obtain the last inequality. Combining the sums over i and j and using the summability assumption on k_{ij}, we have

$$\|v\|_{\mathcal{X}^s}^2 \leq \sum_{(i,j)\in\mathbb{N}\times\mathbb{N}} \left(i^s k_{ij} \right)^2 \sum_{j\in\mathbb{N}} u_j^2 = K^+ \|u\|_{\ell^2}^2. \tag{13.9}$$

This, in turn, implies that indeed $K \in \mathcal{L}(\ell^2; \ell^2)$:

$$\sup_{u\in\ell^2} \frac{\|Ku\|_{\ell^2}}{\|u\|_{\ell^2}} \leq \sup_{u\in\ell^2} \frac{\|Ku\|_{\mathcal{X}^s}}{\|u\|_{\ell^2}} \leq \sqrt{K^+} < \infty.$$

Equation (13.9) also shows that $K \in \mathcal{L}(\ell^2; \mathcal{X}^s)$, i.e., it is bounded when viewed as an operator from ℓ^2 to $\mathcal{X}^s$. Since $\mathcal{X}^s$ is compactly embedded in ℓ^2, we conclude that K is compact as an operator from ℓ^2 to ℓ^2. Thus, we have established that all the conditions of Theorem 13.12 are satisfied, and so we can write

$$K = \sum_{j=1}^{\infty} \lambda_j \varphi^{(j)} \otimes \varphi^{(j)},$$

where $K\varphi^{(j)} = \lambda_j \varphi^{(j)}$.

Now let $e^{(\ell)}$ denote the vector in H with j^{th} entry $\delta_{j\ell}$. Note that $\{e^{(n)}\}_{n\in\mathbb{N}}$ form an orthonormal basis for H. Application of K to $e^{(\ell)}$ delivers the vector with i^{th} entry $k_{i\ell}$. Thus, the Hilbert-Schmidt norm of K is

$$\|K\|_{HS} = \Big(\sum_{(i,\ell)\in\mathbb{N}\times\mathbb{N}} |k_{i\ell}|^2 \Big)^{\frac{1}{2}} \leq \Big(\sum_{(i,\ell)\in\mathbb{N}\times\mathbb{N}} |i^s k_{i\ell}|^2 \Big)^{\frac{1}{2}} < \infty$$

since $s > 0$. Hence we deduce that

$$K_r = \sum_{j=1}^{r} \lambda_j \varphi^{(j)} \otimes \varphi^{(j)}$$

is a finite rank approximation of K which converges to K as $r \to \infty$. $\Diamond$

Exercises

13.1 Let $\|\cdot\|$ denote the Euclidean norm on $\mathbb{R}^m$ and $\mathbb{R}^n$, and the induced operator norm on $\mathcal{L}(\mathbb{R}^p, \mathbb{R}^q)$ with p, q being m or n.

(a) Let $A \in \mathbb{R}^{m\times n}$. Prove that $\|A\| = \sqrt{\rho(A^*A)}$, where ρ denotes the spectral radius: $\rho(M) := \max_j |\lambda_j(M)|$, where $\{\lambda_j(M)\}$ are the eigenvalues of square matrix M.

(b) Let $P \in \mathbb{R}^{n\times n}$ be an orthogonal matrix; using the previous result, show that $\|P\| = 1$. Show also that $\|Pu\| = \|u\|$ for any orthogonal transformation.

(c) Let $U \in \mathbb{R}^{m\times m}$ and $V \in \mathbb{R}^{n\times n}$ be orthogonal matrices and A as in part (a). Show that $\|UAV\| = \|A\|$.

13.2 Let $k\colon [0,1]\times[0,1] \to \mathbb{R}$ be a function such that

(i) $k(t,s) = k(s,t)$ for all $t, s \in [0,1]$;

(ii) $k \in L^2([0,1]\times[0,1])$;

(iii) for any $u \in L^2([0,1])$,

$$\int_0^1 \int_0^1 k(t,s)u(t)u(s)\,dt\,ds \geq 0.$$

Define the operator $K\colon L^2([0,1]) \to L^2([0,1])$ by

$$(Ku)(t) = \int_0^1 k(t,s)u(s)\,ds.$$

(a) Show that K is a symmetric, non-negative, Hilbert-Schmidt operator with

$$\|K\|_{HS} = \|k\|_{L^2([0,1]\times[0,1])}.$$

You may use that if $\{\varphi^{(j)}\}$ is a basis for $L^2([0,1])$, then the tensor product family $\{\psi^{(i,j)}\}$ defined by $\psi^{(i,j)}(t,s) = \varphi^{(i)}(t)\varphi^{(j)}(s)$ is an orthonormal basis for $L^2([0,1]\times[0,1])$.

(b) Show that there exists an orthonormal basis of eigenvectors $\{\varphi^{(j)}\}$ of K with eigenvalues $\{\lambda_j\}$, and that the kernel k may be decomposed as

$$k(t,s) = \sum_{j=1}^{\infty} \lambda_j \varphi^{(j)}(t)\varphi^{(j)}(s).$$

(c) Show that if k is continuous, condition (iii) above may be replaced by the condition that for any $n \in \mathbb{N}$, and any sequences

$t_1, \ldots, t_n \in [0,1]$ and $c_1, \ldots, c_n \in \mathbb{R}$,

$$\sum_{i,j=1}^{n} c_i c_j k(t_i, t_j) \geq 0.$$

13.3 Let

$$H = L^2((0,1);\mathbb{R}) := \left\{ u : (0,1) \to \mathbb{R} \,\middle|\, \int_0^1 u(x)^2 dx < \infty \right\}$$

with inner product

$$\langle u, v \rangle = \int_0^1 u(x)v(x)dx.$$

Consider the differential equation

$$-\frac{d^2u}{dx^2}(x) = r(x), \quad x \in (0,1)$$
$$u(x) = 0, \quad x \in \{0,1\}.$$

(a) Formulate the problem using the Lax-Milgram Theorem 12.12. Prove that there is a unique weak solution u for every $r \in H$.

(b) By using the appropriate Green's function, write the solution in the form $u = Tr$ where T is an integral operator; using this expression show that $T \in \mathcal{L}(H,H)$ and is symmetric.

(c) By using the Fourier sine basis $\{\varphi^{(j)}\}_{j\in\mathbb{N}}$, $\varphi^{(j)} = \sqrt{2}\sin(j\pi x)$, for H, find a representation for T in ℓ^2. Using this, demonstrate that T is compact.

(d) Find the eigenfunctions and eigenvalues of T. Use the approximation theory detailed in the last section to determine a finite rank approximation T_ℓ of T.

(e) Show that compactness alone is enough to establish that T_ℓ is Cauchy in $\mathcal{L}(H,H)$. You may assume that the eigenvalues of a compact operator, ordered to be decreasing in magnitude, tend to zero.

(f) Prove that the eigenvalues of a compact symmetric operator, ordered to be decreasing in magnitude, tend to zero.

13.4(a) Show that $\mathcal{X}^s$ is compactly embedded in $\mathcal{X}^t$ for any $s > t \geq 0$.

(b) Consider operator K defined in Example 13.20. Show that $K \in \mathcal{L}(\mathcal{X}^t;\mathcal{X}^s)$ for any $s > t \geq 0$. What does this tell you about compactness of the operator K viewed as an element of $\mathcal{L}(\mathcal{X}^t;\mathcal{X}^t)$?

13.5 Show that the Definition 13.16 of Hilbert-Schmidt norm of an operator is indeed independent of the choice of orthonormal basis used as stated in Lemma 13.17.

13.6 Show that the finite rank operators $\{A_r\}_{r\in\mathbb{N}}$ form a Cauchy sequence in the Hilbert-Schmidt norm.

14
Singular Value Decomposition

In this lecture we prove the singular value decomposition (SVD) in finite dimensions and then state and prove the analogous result for compact operators in an infinite-dimensional space. These results are the analogue for non-symmetric operators of those in the previous chapter, which was devoted to the symmetric case. As in the previous chapter we also use these ideas as the basis for low rank approximation of the operator.

14.1 Finite Dimensions

Let $A \in \mathbb{R}^{m\times n}$, $A^\top = (a^{(1)}, \dots, a^{(m)})$, $a^{(j)} \in \mathbb{R}^n$ and $\|\cdot\|$ denote Euclidean norm.

Definition 14.1. *Define the* right *and* left singular vectors $\{v^{(j)}, u^{(j)}\}_{j=1}^r$, singular values $\{\sigma_j\}_{j=1}^r$ *and* rank r *as follows:*

$$
\begin{aligned}
v^{(1)} &= \operatorname*{arg\,max}_{\|v\|=1} \|Av\|, \qquad \sigma_1 = \|Av^{(1)}\|, \quad u^{(1)} = \frac{Av^{(1)}}{\sigma_1} \\
v^{(2)} &= \operatorname*{arg\,max}_{\|v\|=1, v\in H_1^\perp} \|Av\|, \quad \sigma_2 = \|Av^{(2)}\|, \quad u^{(2)} = \frac{Av^{(2)}}{\sigma_2} \\
v^{(3)} &= \operatorname*{arg\,max}_{\|v\|=1, v\in H_2^\perp} \|Av\|, \quad \sigma_3 = \|Av^{(3)}\|, \quad u^{(3)} = \frac{Av^{(3)}}{\sigma_3} \\
&\vdots \\
v^{(r)} &= \operatorname*{arg\,max}_{\|v\|=1, v\in H_{r-1}^\perp} \|Av\|, \quad \sigma_r = \|Av^{(r)}\|, \quad u^{(r)} = \frac{Av^{(r)}}{\sigma_r}
\end{aligned}
$$

where $H_1 = \operatorname{span}\{v^{(1)}\}$, $H_2 = \operatorname{span}\{v^{(1)}, v^{(2)}\}$ *and so on, stopping at* r *such that* $\sigma_{r+1} = 0$.

Remark 14.2. $H_j^\perp$ *is closed.* $H_j^\perp \cap \{v : \|v\| = 1\}$ *is compact, hence the* $\arg\max$ *of* $\|Av\|$, *which is sought over a compact set, is attained in* $H_j^\perp \cap \{v : \|v\| = 1\}$.

Definition 14.3. *Let* W *be a* k*-dimensional subspace of* $\mathbb{R}^n$. *Let* $\{w^{(j)}\}_{j=1}^k$ *be an orthonormal basis for* W. *Define, for* $a \in \mathbb{R}^n$,

$$d(a, W) = \|a\|^2 - \sum_{j=1}^{k} |\langle a, w^{(j)}\rangle|^2,$$

the squared distance *of a from* W.

Proposition 14.4. $H_k = \arg\min_{W \in \mathcal{A}_k} \Phi_m(W)$, *where*

$$\mathcal{A}_k = \{W : W \text{ is a } k\text{-dimensional subspace of } \mathbb{R}^n\}$$

and

$$\Phi_m(W) = \sum_{l=1}^{m} d(a^{(l)}, W).$$

Proof Note

$$\Phi_m(W) = \|A\|_F^2 - \sum_{j=1}^{k} \|Aw^{(j)}\|^2.$$

Thus it is equivalent to prove that

$$H_k = \arg\max_{W \in \mathcal{A}_k} \sum_{j=1}^{k} \|Aw^{(j)}\|^2.$$

Let $\{w^{(j)}\}$ be an orthonormal basis for the maximizing set. Since this set has dimension k, we may use Gram-Schmidt to determine this orthonormal basis, choosing $w^{(k)} \perp v^{(1)}, v^{(2)}, \ldots, v^{(k-1)}$ as the first vector in the Gram-Schmidt process and defining the remainder recursively. We now proceed by induction, assuming that the set

$$H_{k-1} = \operatorname{span}\{v^{(1)}, \ldots, v^{(k-1)}\}$$

maximizes over $k-1$ dimensional subspaces so that, in particular,

$$\sum_{j=1}^{k-1} \|Av^{(j)}\|^2 \geq \sum_{j=1}^{k-1} \|Aw^{(j)}\|^2.$$

Now note that $\|Av^{(k)}\|^2 \geq \|Aw^{(k)}\|^2$ by construction of $v^{(k)}$. Thus

$$\sum_{j=1}^{k} \|Av^{(j)}\|^2 \geq \sum_{j=1}^{k} \|Aw^{(j)}\|^2,$$

proving that H_k maximizes $\sum_{j=1}^{k} \|Aw^{(j)}\|^2$ over $\mathcal{A}_k$, and hence that H_k minimizes $\Phi_m(W)$ over $\mathcal{A}_k$ as required. □

Corollary 14.5.

$$\|A\|_F^2 = \sum_{j=1}^{r} \sigma_j^2.$$

Proof The integer r is defined so that $\{v^{(j)}\}_{j=1}^{r}$ spans the row space of A. Thus for $j = 1, \ldots, m$:

$$d(a^{(j)}, H_r) = \|a^{(j)}\|^2 - \sum_{\ell=1}^{r} |\langle a^{(j)}, v^{(\ell)} \rangle|^2$$

so

$$\begin{aligned} 0 &= \sum_{j=1}^{m} d(a^{(j)}, H_r) \\ &= \|A\|_F^2 - \sum_{\ell=1}^{r} \|Av^{(\ell)}\|^2 \\ &= \|A\|_F^2 - \sum_{\ell=1}^{r} \sigma_\ell^2. \end{aligned}$$

□

We have shown:

Theorem 14.6 (Singular Value Decomposition in Finite Dimensions)**.** *Any matrix $A \in \mathbb{R}^{n \times n}$ may be expressed in terms of the right and left singular vectors, the singular values and rank as*

$$A = \sum_{j=1}^{r} \sigma_j u^{(j)} \otimes v^{(j)}.$$

Proof Let $\tilde{A} = \sum_{j=1}^{r} \sigma_j u^{(j)} \otimes v^{(j)}$. We will show that $\tilde{A}v = Av$ for all $v \in \mathbb{R}^n$. Extend $\{v^{(j)}\}_{j=1}^{r}$ to $\{v^{(j)}\}_{j=1}^{n}$ an orthonormal basis for $\mathbb{R}^n$. Necessarily

$Av^{(j)} = 0$ for $j = r+1, \ldots, n$. Now, for $k = 1, \ldots, r$,

$$\tilde{A}v^{(k)} = \sum_{j=1}^{r} \sigma_j u^{(j)} \underbrace{\langle v^{(j)}, v^{(k)} \rangle}_{\delta_{jk}} = \sigma_k u^{(k)} = Av^{(k)}.$$

Similarly $\tilde{A}v^{(k)} = 0$ for $k = r+1, \ldots, n$. Write arbitrary $v \in \mathbb{R}^n$ as

$$v = \sum_{j=1}^{n} \alpha_j v^{(j)}.$$

Then

$$\begin{aligned}
\tilde{A}v &= \sum_{j=1}^{r} \alpha_j \tilde{A}v^{(j)} + \sum_{j=r+1}^{n} \alpha_j \tilde{A}v^{(j)} \\
&= \sum_{j=1}^{r} \alpha_j Av^{(j)} + \sum_{j=r+1}^{n} \alpha_j \times 0 \\
&= \sum_{j=1}^{r} \alpha_j Av^{(j)} + \sum_{j=r+1}^{n} \alpha_j Av^{(j)} \\
&= Av.
\end{aligned}$$

□

Proposition 14.7. *The left singular vectors* $\{u^{(j)}\}_{j=1}^{r}$ *are orthonormal.*

Proof Note that $\|u^{(j)}\| = 1$ for $j = 1, \ldots, r$ by construction. We need to show $\langle u^{(j)}, u^{(k)} \rangle = 0$ for $j \neq k$. We will show $\langle u^{(j)}, u^{(1)} \rangle = 0$ for $j = 2, \ldots, r$. Others are proved similarly. Suppose for contraction that there exists $\delta > 0$ such that

$$\langle u^{(j)}, u^{(1)} \rangle = \delta > 0$$

for some $j \in \{2, \ldots, r\}$ (the case $\delta < 0$ can be handled similarly). Define

$$v_*^{(1)} = \frac{v^{(1)} + \epsilon v^{(j)}}{\|v^{(1)} + \epsilon v^{(j)}\|}$$

so that $\|v_*^{(j)}\| = 1$. Then

$$Av_*^{(1)} = \frac{\sigma_1 u^{(1)} + \epsilon \sigma_j u^{(j)}}{(1+\epsilon^2)^{1/2}}.$$

Now, using this identity, for all $\epsilon > 0$ sufficiently small,

$$\begin{aligned}\|Av_*^{(1)}\| &= \|u^{(1)}\|\|Av_*^{(1)}\| \\ &\geq |\langle u^{(1)}, Av_*^{(1)}\rangle| \\ &= \frac{\sigma_1 + \epsilon\delta\sigma_j}{(1+\epsilon^2)^{1/2}} \\ &\geq \sigma_1 + \frac{1}{2}\epsilon\delta\sigma_j \\ &> \sigma_1.\end{aligned}$$

But $\|Av_*^{(1)}\| \leq \|Av^{(1)}\| = \sigma_1$ by definition, thus $\sigma_1 > \sigma_1$, a contradiction and hence $\delta = 0$. □

Recall that in the proof of Theorem 14.6 we extended the right singular vectors to an orthonormal basis for $\mathbb{R}^n$: $\{v^{(j)}\}_{j=1}^n$. We may do the same thing for left singular vectors: $\{u^{(j)}\}_{j=1}^n$; and we may also extend the singular values to a set of cardinality n by defining $\sigma_j = 0$ for $j = r+1, \cdots, n$: $\{\sigma_j\}_{j=1}^n$. Thus we have

$$\begin{aligned}&Av^{(j)} = \sigma_j u^{(j)}, \quad j = 1, \cdots, n \\ &\langle v^{(i)}, v^{(j)}\rangle = \langle u^{(i)}, u^{(j)}\rangle = \delta_{ij}, \quad i, j = 1, \cdots, n.\end{aligned}$$

Creating orthogonal matrices V and U with the $\{v^{(j)}\}_{j=1}^n$ and $\{u^{(j)}\}_{j=1}^n$ as columns respectively, and diagonal matrix Σ with entries $\{\sigma_j\}_{j=1}^n$ we have the following form of the SVD:

Theorem 14.8. *For any matrix $A \in \mathbb{R}^{n\times n}$ there are orthogonal matrices U, V and diagonal matrix Σ with non-negative entries such that $A = U\Sigma V^T$.*

Definition 14.9. *Let $A = U\Sigma V^*$ be an SVD for $A \in \mathbb{C}^{m\times n}$. Then the* Moore–Penrose pseudoinverse *$A^\dagger \in \mathbb{C}^{n\times m}$ of A is defined by $A^\dagger := V\Sigma^\dagger U^*$, where $\Sigma^\dagger \in \mathbb{R}^{n\times m}$ is the transpose of the matrix obtained from $\Sigma \in \mathbb{R}^{m\times n}$ by replacing all strictly positive singular values σ_k by their reciprocals $1/\sigma_k$.*

Remark 14.10. *If $A \in \mathbb{C}^{n\times n}$ is invertible, then $\operatorname{rank}(A) = n$ and it can be shown that $A^\dagger = A^{-1}$. Thus, the Moore–Penrose pseudoinverse is a generalization of the usual matrix inverse to more general matrices.*

14.2 Singular Value Decomposition for Compact Operators

Let $(X, \langle\cdot,\cdot\rangle_X, \|\cdot\|_X)$ and $(Y, \langle\cdot,\cdot\rangle_Y, \|\cdot\|_Y)$ be Hilbert spaces and $K : X \to Y$ a compact operator. We will derive a singular value decomposition for such K. The following lemma will be important in what follows.

Lemma 14.11. *Let $K \in \mathcal{L}(X, Y)$ be a compact operator between the Hilbert spaces X and Y. The operator $K^*K \in \mathcal{L}(X, X)$ is compact and symmetric; furthermore, $\mathrm{Ker}(K^*K)^\perp = \mathrm{Ker}(K)^\perp$. Finally, $\|K^*K\| = \|K\|^2$ in the natural induced operator norms.*

Proof (Sketch) The symmetry of K^*K is inherent in the definition. The compactness result follows from the fact that K^* is bounded since K is bounded, and from the fact that the composition of a bounded operator with a compact operator is compact (we do not prove this in these notes). The fact that $\mathrm{Ker}(K) \subseteq \mathrm{Ker}(K^*K)$ is clear; the fact that their orthogonal complements coincide is not proved in these notes. The final result follows from Lemma 13.11, which delivers the final identity as follows:

$$\begin{aligned}
\|K^*K\| &= \sup_{\|f\|_X=1} |\langle f, K^*Kf\rangle_X| \\
&= \sup_{\|f\|_X=1} |\langle Kf, Kf\rangle_Y| \\
&= \sup_{\|f\|_X=1} \|Kf\|_Y^2 \\
&= \left(\sup_{\|f\|_X=1} \|Kf\|_Y\right)^2 \\
&= \|K\|^2.
\end{aligned}$$

□

To derive the singular value decomposition of K, we proceed as follows. We apply Theorem 13.12 to $A = K^*K$ which, by the preceding lemma, is compact, and by construction is symmetric. We obtain $X-$orthonormal eigenvectors $\{v^{(n)}\}_{n\in\mathbb{N}}$ and eigenvalues $\{\lambda_n\}_{n\in\mathbb{N}}$ satisfying

$$K^*Kv^{(n)} = \lambda_n v^{(n)}.$$

By the preceding lemma, the $\{v^{(n)}\}_{n\in\mathbb{N}}$ are an orthonormal basis for $\mathrm{Ker}(K^*K)^\perp = \mathrm{Ker}(K)^\perp$ and we refer to them as *right singular vectors*. Since K^*K is positive semi-definite, the eigenvalues $\{\lambda_n\}_{n\in\mathbb{N}}$ are all non-negative and we may define $\sigma_n = \sqrt{\lambda_n}$ to be the (real) *singular values*. The

left singular vectors $\{u^{(n)}\}_{n\in\mathbb{N}}$ are then defined by

$$u^{(n)} = \frac{1}{\sigma_n} K v^{(n)}, \quad n \in \mathbb{N}$$

and, by construction, satisfy $Y-$orthonormality:

$$\langle u^{(n)}, u^{(m)} \rangle_Y = \frac{1}{\sigma_n \sigma_m} \langle v^{(n)}, K^* K v^{(m)} \rangle_X = \frac{\lambda_m}{\sigma_n \sigma_m} \delta_{nm} = \delta_{nm}.$$

Furthermore, also by construction,

$$K^* u^{(n)} = \frac{1}{\sigma_n} K^* K v^{(n)} = \sigma_n v^{(n)}, \quad n \in \mathbb{N}.$$

It follows that

$$KK^* u^{(n)} = \sigma_n^2 u^{(n)} = \lambda_n u^{(n)},$$

and thus the left singular vectors are eigenvectors of KK^*. By Theorem 13.12, any $f \in X$ can be written as

$$f = f_0 + \sum_{n=1}^{\infty} \langle v^{(n)}, f \rangle_X v^{(n)},$$

where $f_0 \in \mathrm{Ker}(K)$, using the preceding lemma to note that $\mathrm{Ker}(K^*K)^\perp = \mathrm{Ker}(K)^\perp$. Recalling Definition 13.14, let $g \in \mathrm{Ran}(K)$ and note that $g = Kf$ for any preimage $f \in X$ of g. Thus

$$g = Kf = \sum_{n=1}^{\infty} \langle v^{(n)}, f \rangle_X K v^{(n)} = \sum_{n=1}^{\infty} \sigma_n \langle v^{(n)}, f \rangle_X u^{(n)}.$$

Using this it follows that the *left singular vectors* $\{u^{(n)}\}_{n\in\mathbb{N}}$ form an orthonormal basis for $\overline{\mathrm{Ran}(K)}$. We have proved the following:

Theorem 14.12 (Singular Value Decomposition)**.** *Every compact operator K between Hilbert spaces X and Y can be written in the form*

$$K = \sum_{n=1}^{\infty} \sigma_n u^{(n)} \otimes v^{(n)}, \tag{14.1}$$

where $\|K\| = \sigma_1 \geq \sigma_2 \geq \cdots$ are the (at most countably many) non-negative singular values, and the left and right singular vectors $\{u^{(n)}\}_{n\in\mathbb{N}}$, $\{v^{(n)}\}_{n\in\mathbb{N}}$, as orthonormal bases for $\overline{\mathrm{Ran}(K)}$ and $\mathrm{Ker}(K)^\perp$, respectively, satisfy

$$K v^{(n)} = \sigma_n u^{(n)}, \quad K^* u^{(n)} = \sigma_n v^{(n)}.$$

14.3 Approximation of Compact Operators

For simplicity we consider the setting from the preceding section, but with $X = Y = H$ where $\Big(H, \langle \cdot\,, \cdot \rangle, \|\cdot\|\Big)$ is a Hilbert space. We consider the compact operator $K \in \mathcal{L}(H, H)$ defined there, assuming that $\mathrm{Ker}(K)^\perp$ is infinite-dimensional. In view of formula (14.1), a natural rank r approximation of K is

$$K_r = \sum_{n=1}^{r} \sigma_n u^{(n)} \otimes v^{(n)}. \tag{14.2}$$

Recall the Definition 13.16 of a Hilbert-Schmidt operator (see also Exercise 10.8). Choosing $\{e^{(n)} = v^{(n)}\}_{n\in\mathbb{N}}$ in the definition of the Hilbert-Schmidt norm of K gives

$$\|K\|_{HS} = \Big(\sum_{n=1}^{\infty} \sigma_n^2\Big)^{\frac{1}{2}}.$$

Thus compact $K \in \mathcal{L}(H, H)$ is Hilbert-Schmidt if and only if the ordered singular value sequence is an element of $\ell^2(\mathbb{N}; \mathbb{R})$.

Theorem 14.13. *Let $K \in \mathcal{L}(H, H)$ be compact and Hilbert-Schmidt, and assume that $\mathrm{Ker}(K)^\perp$ is infinite-dimensional. Then the finite rank approximation K_r given by* (14.2) *converges to A given by* (14.1) *in the sense that $\|K - K_r\|_{HS} \to 0$ as $r \to \infty$.*

Proof It is straightforward that

$$K - K_r = \sum_{n=r+1}^{\infty} \sigma_n u^{(n)} \otimes v^{(n)}.$$

Furthermore

$$(K - K_r)v^{(\ell)} = \sum_{n=r+1}^{\infty} \sigma_n u^{(n)} \langle v^{(n)}, v^{(\ell)} \rangle = \sigma_\ell u^{(\ell)},$$

for $\ell \geq r + 1$ and is zero otherwise. Thus

$$\|K - K_r\|_{HS}^2 = \sum_{\ell=r+1}^{\infty} \sigma_\ell^2.$$

Because K is Hilbert-Schmidt, we have

$$\sum_{\ell=1}^{\infty} \sigma_\ell^2 < \infty,$$

from which it follows that

$$\sum_{\ell=r+1}^{\infty} \sigma_\ell^2 \to 0$$

as $r \to \infty$ and the result is proved. □

Exercises

14.1 Find an SVD for each of the following matrices:

$$A = \begin{pmatrix} 2 & 2 \\ -1 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \end{pmatrix}.$$

14.2 Let $A \in \mathbb{C}^{n\times n}$ and let $\sigma_1 \geq \ldots \geq \sigma_n > 0$ be the singular values of A. Show that A is invertible with $\|A^{-1}\|_2 = 1/\sigma_n$.

14.3 Let $A \in \mathbb{C}^{m\times n}$. Using the SVD of A, show that there exists a unitary matrix $S \in \mathbb{C}^{m\times n}$ and a non-negative matrix $P \in \mathbb{C}^{n\times n}$ such that $A = SP$. Show also that there exists a unitary matrix $\tilde{S} \in \mathbb{C}^{n\times m}$ and a non-negative matrix $\tilde{P} \in \mathbb{C}^{m\times m}$ such that $A = \tilde{P}\tilde{S}$. What are these decompositions in the case $n = m = 1$?

14.4 Recall the Definition 10.3 of a Hilbert-Schmidt operator (see also Exercise 10.8). Show that operator K (Eqn. 14.1) from Theorem 14.12 is Hilbert-Schmidt if $\sum_{j\in\mathbb{N}} \sigma_j^2 < \infty$. Deduce that in this setting the finite-rank sequence $K_r \in \mathcal{L}(H, H)$ is Cauchy with respect to the Hilbert-Schmidt norm on the space of bounded linear operators from H into itself.

14.5 Consider the setting of Theorem 14.13. Find approximations for $A := KK^*$ and $B := K^*K$, respectively, involving only the set $\{\sigma_j, u^{(j)}, v^{(j)}\}_{j\in\mathbb{N}}$, which converge to A and B in $\mathcal{L}(H, H)$.

15
Jordan Normal Form

In this lecture we deal with the matrix decomposition known as the *Jordan normal form* (JNF) (sometimes also known as the *Jordan canonical form* (JCF).) Whilst matrix diagonalization is not possible for all matrices, the JNF always exists, for any matrix; in this regard it is similar to the SVD. It should be noted, though, that numerically computing the JNF for a given matrix A is, in general, hard; this is because the JNF itself is not, in general, stable to arbitrarily small perturbations. In contrast the SVD is continuous with respect to perturbations in the matrix in A. Both the JNF and SVD are useful for theoretical purposes, often complementary to one another. The situation is summarized in the following table:

	Algorithms	Theory
SVD	✓	✓
JNF		✓

15.1 Spectral Radius

Definition 15.1. *Let $A \in \mathbb{C}^{n\times n}$. The* spectral radius $\rho(A)$ *of a matrix A is a non-negative real number defined by*

$$\rho(A) \coloneqq \max\left\{|\lambda| : \lambda \text{ is an eigenvalue of } A\right\}.$$

Theorem 15.2. *For any induced matrix norm $\|\cdot\|$ on $\mathbb{C}^{n\times n}$, any $A \in \mathbb{C}^{n\times n}$ and any $k \in \mathbb{N}$,*

$$\rho(A)^k \leq \rho(A^k) \leq \|A^k\| \leq \|A\|^k.$$

Proof For the first inequality, we notice that if λ is an eigenvalue of A, then λ^k is an eigenvalue of A^k: if $A\varphi = \lambda\varphi$, with $\varphi \neq 0$ (by definition of an eigenvector), then $A^k\varphi = \lambda^k\varphi$. We then have:

$$\begin{aligned}\rho(A)^k &= \max\left\{|\lambda|^k : \lambda \text{ is an eigenvalue of } A\right\}\\ &\leq \max\left\{|\gamma| : \gamma \text{ is an eigenvalue of } A^k\right\}\\ &= \rho(A^k).\end{aligned}$$

Now let $B = A^k$. By definition of the spectral radius of B, there exists non-zero $x \in \mathbb{C}^n$ such that

$$Bx = \gamma x, \quad |\gamma| = \rho(B).$$

Let $X = (x, \ldots, x) \in \mathbb{C}^{n\times n}$, then $BX = \gamma X$. Using the fact that the norm is induced and, hence, submultiplicative, we have

$$\|B\|\|X\| \geq \|BX\| = \|\gamma X\| = |\gamma|\|X\| = \rho(B)\|X\|.$$

Therefore, $\rho(B) \leq \|B\|$, which implies the second inequality of the theorem.

Finally, $\|A^k\| = \|AA^{k-1}\| \leq \|A\|\,\|A^{k-1}\|$, which, by induction, leads to the last bound. □

Theorem 15.3. *If $A \in \mathbb{C}^{n\times n}$ has n orthogonal eigenvectors, then*

$$\rho(A)^k = \rho(A^k) = \|A^k\|_2 = \|A\|_2^k$$

for any $k \in \mathbb{N}$.

Proof Let $\left\{x^{(j)}\right\}_{j=1}^n$ be an orthonormal basis composed of eigenvectors of A with corresponding eigenvalues $\left\{\lambda_j\right\}_{j=1}^n$. Without loss of generality, take $\rho(A) = |\lambda_1|$, thus $|\lambda_1| \geq |\lambda_j|$ for any $j = 2, \ldots, n$.

Let $x \in \mathbb{C}^n$ be arbitrary and express it in the basis as

$$x = \sum_{j=1}^n \alpha_j x^{(j)}, \quad \alpha_j = \langle x^{(j)}, x\rangle.$$

Then we have

$$Ax = \sum_{j=1}^n \alpha_j\lambda_j x^{(j)},$$

$$\|Ax\|_2^2 = \sum_{j=1}^n |\alpha_j\lambda_j|^2 \leq |\lambda_1|^2 \sum_{j=1}^n |\alpha_j|^2 = |\lambda_1|^2\|x\|_2^2,$$

where in the last step we used the Pythagorean theorem. This shows that

$$\frac{\|Ax\|_2}{\|x\|_2} \leq |\lambda_1| = \rho(A)$$

for all $x \neq 0$ and consequently

$$\|A\|_2 = \max_{x \neq 0} \frac{\|Ax\|_2}{\|x\|_2} \leq \rho(A).$$

Combining this result with the previous theorem, we conclude that $\rho(A) = \|A\|_2$. Thus, for all $k \in \mathbb{N}$, we have

$$\begin{aligned} \|A\|_2^k = \rho(A)^k \leq \rho(A^k) \leq \|A^k\|_2 \leq \|A\|_2^k \quad \Rightarrow \\ \rho(A)^k = \rho(A^k) = \|A^k\|_2 = \|A\|_2^k. \end{aligned}$$

□

A general matrix $A \in \mathbb{C}^{n \times n}$ does not have n orthogonal eigenvectors. Obtaining a relationship between the norm and spectral radius of A requires consideration of the spectral radius of A^*A:

Theorem 15.4. *For any $A \in \mathbb{C}^{n \times n}$, it follows that $\|A\|_2^2 = \rho(A^*A)$.*

Proof See Exercise 15.5. □

15.2 Jordan Normal Form

Given the two preceding theorems, it is natural to ask whether there is a general relationship between some norm of A and its spectral radius, outside of the "n orthogonal eigenvectors" setting. The answer is "nearly yes" and is contained in Theorem 15.11; see also Remark 15.12. To prove it we make use of the JNF which we first state.

Definition 15.5. *For $\lambda \in \mathbb{C}$, $n \in \mathbb{N}$, a* Jordan block *$J_n(\lambda) \in \mathbb{C}^{n \times n}$ is a matrix satisfying*

$$\{J_n(\lambda)\}_{ij} = \begin{cases} \lambda, & i = j \quad \textit{(on diagonal)} \\ 1, & i = j - 1 \quad \textit{(on super-diagonal)} \\ 0, & \textit{otherwise}. \end{cases}$$

Example 15.6. For $n = 3$, a Jordan block $J_3(\lambda)$ looks like this:

$$\begin{pmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{pmatrix}.$$

◊

Definition 15.7. *A* Jordan matrix $J \in \mathbb{C}^{n\times n}$ *is a block-diagonal matrix of the following form:*

$$J = \begin{pmatrix} J_{n_1}(\lambda_1) & & & \\ & J_{n_2}(\lambda_2) & & 0 \\ 0 & & \ddots & \\ & & & J_{n_k}(\lambda_k) \end{pmatrix}, \tag{15.1}$$

where $\sum_{j=1}^{k} n_j = n$ *and* $\{\lambda_j\}_{j=1}^{k} \subset \mathbb{C}$, $k \leq n$, $n_j \geq 1$.

Theorem 15.8 (Jordan Canonical Form)**.** *For any* $A \in \mathbb{C}^{n\times n}$*, there exist an invertible matrix* $S \in \mathbb{C}^{n\times n}$ *and a Jordan matrix* $J \in \mathbb{C}^{n\times n}$ *such that* $A = SJS^{-1}$*. The diagonal elements* $\left\{\lambda_j\right\}_{j=1}^{k}$ *implicit in* J *are the eigenvalues of* A*.*

The theorem here is stated without a proof.

Remark 15.9. *Even for* $A \in \mathbb{R}^{n\times n}$*, the JNF may be complex. This is easily seen noting that the eigenvalues of a real-valued matrix are, in general, complex numbers.*

Lemma 15.10. *Let* $D_n^\delta = \operatorname{diag}\{\delta, \delta^2, \ldots, \delta^n\}$ *for any* $\delta > 0$*. Then*

$$\left(D_n^\delta\right)^{-1} J_n(\lambda) D_n^\delta = J_n^\delta(\lambda),$$

where

$$\left\{J_n^\delta(\lambda)\right\}_{ij} = \begin{cases} \lambda, & i = j \quad \textit{(on diagonal)} \\ \delta, & i = j-1 \quad \textit{(on super-diagonal)} \\ 0, & \textit{otherwise}. \end{cases}$$

Proof The proof can be carried out using induction. Here we only provide an example for $n = 3$, with the inductive step being very similar to the one in the proof of the finite-dimensional Spectral Theorem in Lecture 13.

For $n = 3$, we write

$$\begin{aligned}\left(D_n^\delta\right)^{-1} J_n(\lambda) D_n^\delta &= \begin{pmatrix}\delta^{-1} & 0 & 0\\ 0 & \delta^{-2} & 0\\ 0 & 0 & \delta^{-3}\end{pmatrix}\begin{pmatrix}\lambda & 1 & 0\\ 0 & \lambda & 1\\ 0 & 0 & \lambda\end{pmatrix}\begin{pmatrix}\delta^1 & 0 & 0\\ 0 & \delta^2 & 0\\ 0 & 0 & \delta^3\end{pmatrix}\\ &= \begin{pmatrix}\delta^{-1} & 0 & 0\\ 0 & \delta^{-2} & 0\\ 0 & 0 & \delta^{-3}\end{pmatrix}\begin{pmatrix}\delta\lambda & \delta^2 & 0\\ 0 & \delta^2\lambda & \delta^3\\ 0 & 0 & \delta^3\lambda\end{pmatrix}\\ &= \begin{pmatrix}\lambda & \delta & 0\\ 0 & \lambda & \delta\\ 0 & 0 & \lambda\end{pmatrix} = J_n^\delta(\lambda).\end{aligned}$$

□

Theorem 15.11. *Let $A \in \mathbb{C}^{n\times n}$ and $\delta > 0$. Then there exists a norm $\|\cdot\|$ on $\mathbb{C}^n$ such that the induced matrix norm satisfies*

$$\rho(A) \le \|A\| \le \rho(A) + \delta.$$

Proof The lower bound is true for all matrix norms; see Theorem 15.2. For the upper bound, define matrix D^δ by

$$D^\delta = \begin{pmatrix} D_{n_1}^\delta & & 0\\ & \ddots & \\ 0 & & D_{n_k}^\delta\end{pmatrix}.$$

Since D^δ, J and $\left(D^\delta\right)^{-1}$ are block-diagonal matrices with the same respective block sizes, we have, according to the previous lemma,

$$\begin{aligned}\left(D^\delta\right)^{-1} J D^\delta &= J^\delta,\\ J &= D^\delta J^\delta \left(D^\delta\right)^{-1},\end{aligned}$$

where

$$J^\delta = \begin{pmatrix} J_{n_1}^\delta(\lambda_1) & & & 0\\ & J_{n_2}^\delta(\lambda_2) & & \\ 0 & & \ddots & \\ & & & J_{n_k}^\delta(\lambda_k)\end{pmatrix}.$$

Define the norm on $\mathbb{C}^n$ by $\|x\| = \|(SD^\delta)^{-1}x\|_\infty$, where S is the matrix given by JNF. Then we can write:

$$A = SJS^{-1} = \left(SD^\delta\right) J^\delta \left(SD^\delta\right)^{-1}.$$

Thus,

$$\begin{aligned}\|A\| &= \max_{x\neq 0}\frac{\|Ax\|}{\|x\|} = \max_{x\neq 0}\frac{\left\|\left(SD^{\delta}\right)^{-1}Ax\right\|_{\infty}}{\left\|\left(SD^{\delta}\right)^{-1}x\right\|_{\infty}}\\ &= \max_{x\neq 0}\frac{\left\|J^{\delta}\left(SD^{\delta}\right)^{-1}x\right\|_{\infty}}{\left\|\left(SD^{\delta}\right)^{-1}x\right\|_{\infty}}\\ &= \max_{y\neq 0}\frac{\left\|J^{\delta}y\right\|_{\infty}}{\|y\|_{\infty}} \qquad \text{(since } SD^{\delta} \text{ is invertible)}\\ &= \left\|J^{\delta}\right\|_{\infty}.\end{aligned}$$

From the definition of the $\|\cdot\|_{\infty}$ norm, we see that

$$\left\|J^{\delta}\right\|_{\infty} \leq \max_{1\leq j\leq k}|\lambda_j| + \delta = \rho(A) + \delta.$$

Thus $\|A\| \leq \rho(A) + \delta$. This concludes the proof.

□

Remark 15.12. *If $n_j = 1$ for all $j = 1, \ldots, k$, which implies $k = n$, then careful study of the preceding theorem shows that, since $J_{n_j}(\lambda_j) = \lambda_j$, there exists a norm such that $\|A\| = \rho(A)$.*

15.3 Matrix Functions

Now we proceed to extend scalar functions to functions acting on linear bounded operators in a Banach space and, in particular, extensions of scalar functions to matrix functions. The latter will use the JNF. The construction through powers of operators rests on the Banach algebra properties introduced in Lecture 7.

Let $f : \mathbb{C} \to \mathbb{C}$ with

$$f(z) = \sum_{j=0}^{\infty} a_j z^j, \quad \sum_{j=0}^{\infty} |a_j||z|^j < \infty \quad \forall |z| < r,$$

where $\{a_j\}_{j\in\mathbb{Z}^+} \subset \mathbb{C}$. Now let $(X, \|\cdot\|)$ be Banach over $\mathbb{K} = \mathbb{C}$; for $A \in \mathcal{L}(X, X)$, define $f(A)$ by

$$f(A) = \lim_{n\to\infty} f_n(A), \quad f_n(A) = \sum_{j=0}^{n} a_j A^j.$$

Theorem 15.13. *For each $A \in \mathcal{L}(X,X)$ with $\|A\| < r$, the operator sequence f_n converges to f in $\mathcal{L}(X,X)$: $\|f_n - f\|_{\mathcal{L}(X,X)} \to 0$ as $n \to \infty$.*

Proof According to the theory developed in Lecture 7, it suffices to prove that $\{f_n\}$ is Cauchy in $\mathcal{L}(X,X)$ (X is Banach $\Rightarrow$ $\mathcal{L}(X,X)$ is Banach).

Clearly, $f_n \in \mathcal{L}(X,X)$ for any $n \in \mathbb{N}$, since $A \in \mathcal{L}(X,X)$. Without loss of generality, let $n > m$, $n, m \in \mathbb{N}$, then

$$\|f_n(A) - f_m(A)\| = \left\| \sum_{j=m+1}^{n} a_j A^j \right\| \leq \sum_{j=m+1}^{n} |a_j| \|A\|^j.$$

From the requirements on f and the fact that $\|A\| < r$, we know that the series above converges and, therefore, is Cauchy in $\mathbb{R}$. By choosing $n, m \geq N = N(\varepsilon)$, we have

$$\|f_n(A) - f_m(A)\| < \varepsilon \quad \forall n, m \geq N(\varepsilon).$$

Hence f_n is Cauchy in $\mathcal{L}(X,X)$. □

Remark 15.14. *Let $A \in \mathbb{C}^{n\times n}$ as before.*

$$A = SJS^{-1} \;\Rightarrow\; A^j = SJ^jS^{-1}.$$

Thus if $f(A) = \sum_{j=0}^{\infty} a_j A^j$, then

$$f(A) = S\left(\sum_{j=0}^{\infty} a_j J^j\right) S^{-1} = Sf(J)S^{-1}.$$

In other words, we only need to make sense of the action of a function f on Jordan matrices, $f(J)$. Moreover, if J has a block-diagonal structure as in (15.1)*, then*

$$J^j = \begin{pmatrix} J_{n_1}(\lambda_1)^j & & & \\ & J_{n_2}(\lambda_2)^j & & 0 \\ & & \ddots & \\ 0 & & & J_{n_k}(\lambda_k)^j \end{pmatrix},$$

$$f(J) = \begin{pmatrix} f\left(J_{n_1}(\lambda_1)\right) & & & \\ & f\left(J_{n_2}(\lambda_2)\right) & & 0 \\ & & \ddots & \\ 0 & & & f\left(J_{n_k}(\lambda_k)\right) \end{pmatrix}.$$

Thus the problem of defining matrix functions further reduces to defining

$f(J_n(\lambda))$: the action of function f on an arbitrary Jordan block. This will allow us to define $f(A)$ for any $A \in \mathbb{C}^{n\times n}$. It is how we start off the next lecture.

Exercises

15.1 *For this exercise, use either* `EigTool` *(for MATLAB):*

http://github.com/eigtool/eigtool

or `PseudoPy` *(for Python):*

https://github.com/andrenarchy/pseudopy

or any other package of your choice to compute pseudospectra. See remark about using `EigTool` *at the end of this exercise.*

We will work with the 2-norm unless stated otherwise in this exercise.
Given a matrix $A \in \mathbb{C}^{n\times n}$, we define its spectrum $\Lambda(A)$ by

$$\Lambda(A) = \{\lambda \in \mathbb{C} : \lambda I - A \text{ is not invertible}\}.$$

That is, $\Lambda(A)$ is the set of eigenvalues of A. Throughout this exercise and Exercises 15.3 and 15.4, we will study the ε-pseudospectrum of A, denoted $\Lambda_\varepsilon(A)$. This is a subset of the complex plane which contains $\Lambda(A)$, and in some sense can be thought of as the set of "approximate eigenvalues" of A.

(a) Look up at least two equivalent definitions of the ε-pseudospectrum of a matrix online and state them.

(b) Let $A \in \mathbb{C}^{n\times n}$, $z \in \mathbb{C}$ and $\varepsilon > 0$. Show that the following are equivalent:

(i) $\|(zI - A)^{-1}\|_2 \geq \varepsilon^{-1}$;

(ii) z is an eigenvalue of $A + E$ for some $E \in \mathbb{C}^{n\times n}$ with $\|E\|_2 \leq \varepsilon$;

(iii) there exists a vector $u \in \mathbb{C}^n$ with $\|u\|_2 = 1$ and $\|(zI-A)u\|_2 \leq \varepsilon$;

(iv) $\sigma_{\min}(zI - A) \leq \varepsilon$, where $\sigma_{\min}(zI - A)$ is the smallest singular value of $zI - A$.

Hint: Suggested order of implications is (i) $\Rightarrow$ (iii) $\Rightarrow$ (ii) $\Rightarrow$ (i) and (i) $\Leftrightarrow$ (iv).

(c) Show that when A is normal, i.e., $A^*A = AA^*$, its ε-pseudospectrum is given by

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} : \operatorname{dist}(z, \Lambda(A)) \leq \varepsilon\}$$

where $\operatorname{dist}(z, \Lambda(A)) = \inf_{\lambda \in \Lambda(A)} |z - \lambda|$ is the distance from the point z to the set $\Lambda(A)$. Describe how this pseudospectrum looks geometrically in the complex plane.

(d) Define $A \in \mathbb{C}^{2\times 2}$ by

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}.$$

Show that A is not normal. Find an explicit function $f : \mathbb{C} \to \mathbb{R}$ such that[1]

$$\Lambda_\varepsilon(A) = \{z \in \mathbb{C} \,|\, f(z) \leq \varepsilon\}.$$

Using a programming language of your choice, plot the contours of f for a variety of levels $\varepsilon \in (0, 1)$. Contrast this with what would be expected if A was normal.

Remark 15.15. *To use EigTool on a matrix* `A`*, call* `eigtool(A)` *from the directory where you have extracted the software. To customize options for plotting, in the window that opens, select Extras ▸ Options ▸ Code for Printing. Copy and paste the code for the struct* `opts` *that appears in the console, changing it where appropriate, and then call* `eigtool(A,opts)`*. In particular, increasing the value of* `opts.npts` *and the number of levels in* `opts.levels` *should produce better plots.*

15.2 Here we study properties of matrices that arise in the theory of Markov chains. Let $S = \{s_1, \ldots, s_n\}$ denote a finite state space. A Markov chain on S is a discrete time stochastic process $\{X_k\}$ taking values in S such that its future trajectory depends only upon its current state. In this exercise we are only concerned with time homogeneous Markov chains, so that for any time $k \in \mathbb{N}$ and any states $s_i, s_j \in S$,

$$\mathbb{P}(X_{k+1} = s_j \,|\, X_k = s_i) = \mathbb{P}(X_1 = s_j \,|\, X_0 = s_i).$$

The Markov chain can then be completely characterized by its one-step transition probabilities.

[1] Do not worry about simplifying f.

Denote by P_{ij} the probability of moving from state i to state j in a single step:

$$P_{ij} = \mathbb{P}(X_1 = s_j \mid X_0 = s_i).$$

The matrix $P = \{P_{ij}\}$ is called the transition matrix for $\{X_k\}$.

We will say that $\nu \in \mathbb{R}^n$ is a distribution if $\nu_j \in [0, 1]$ for all j and $\|\nu\|_1 = 1$. We will say that a random variable U on S is distributed according to ν if for each j,

$$\mathbb{P}(U = s_j) = \nu_j.$$

Given a Markov chain $\{X_k\}$ with transition matrix P, a distribution μ is said to be an invariant distribution for $\{X_k\}$ if

$$P^*\mu = \mu.$$

Thus μ is a right eigenvector of P^* with eigenvalue 1, or equivalently μ^* is a left eigenvector of P with eigenvalue 1. Let $F\colon S \to \mathbb{R}$ be a function on S. Then associated with F is a vector $f \in \mathbb{R}^n$ given by $f_j = F(s_j)$ for each j. The expected value of F under μ can then be computed as

$$\mathbb{E}\left(F(U)\right) \coloneqq \sum_{j=1}^{n} F(s_j)\,\mathbb{P}(U = s_j) = \sum_{j=1}^{n} f_j \mu_j = \langle f, \mu \rangle_2$$

where U is distributed according to μ.

Given a Markov chain $\{X_k\}$ with an invariant distribution μ, the question arises of whether the chain converges to μ in a distributional sense. If this is the case, $\{X_k\}$ is said to be *ergodic*. Formally this means that for large enough k, the statistics of the chain coincide with those of samples from μ. When this is the case, we may approximate

$$\langle f, \mu \rangle_2 \approx \frac{1}{N} \sum_{k=1}^{N} F(X_k)$$

without the need to know or compute μ. This can be interpreted as approximating the spatial average $\mathbb{E}\left(F(U)\right)$ by the time average of $F(X_k)$.

We will give sufficient conditions for $\{X_k\}$ to be ergodic, but first we must introduce the notions of irreducibility and aperiodicity:

Definition 15.16. *Let $\{X_k\}$ be a Markov chain on S with transition matrix P.*

(i) $\{X_k\}$ *is said to be* irreducible *if there is a positive probability of moving between any two states* s_i, s_j *in a finite time.*

(ii) $\{X_k\}$ *is said to be* aperiodic *if* $\gcd\{m \geq 1 : (P^m)_{jj} > 0\} = 1$ *for each* j.

Theorem 15.17. *Let* $\{X_k\}$ *be an irreducible aperiodic Markov chain on* S *with transition matrix* P *and invariant distribution* μ. *Then, we have*

$$P^\infty := \lim_{m\to\infty} P^m = \begin{pmatrix} \mu_1 & \mu_2 & \dots & \mu_n \\ \mu_1 & \mu_2 & \dots & \mu_n \\ \vdots & \vdots & \ddots & \vdots \\ \mu_1 & \mu_2 & \dots & \mu_n \end{pmatrix}.$$

Note that in particular the above result implies that the invariant distribution is unique. If the hypotheses of the theorem are satisfied, it hence makes sense to refer to *the* invariant distribution for $\{X_k\}$.

(a) Show that the m-step transition probabilities are given by

$$\mathbb{P}(X_m = s_j \mid X_0 = s_i) = (P^m)_{ij}.$$

(b) Show that the left and right eigenvalues of a matrix coincide. By finding a left eigenvector for P^*, deduce that an invariant distribution for $\{X_k\}$ always exists. (Potential pitfall: remember that for $\xi \in \mathbb{R}^n$ to be a distribution, it should hold that $\xi_i \in [0, 1]$ and $\|\xi\|_1 = 1$.)

(c) Let $\{X_k\}$ satisfy the hypotheses of Theorem 15.17.

(i) Show that given any distribution ν, we have $(P^m)^*\nu \to \mu$.

(ii) Define $A = P - P^\infty$. Show that $A^m = P^m - P^\infty$ for all $m \geq 1$.

15.3 In this exercise, we introduce the Ehrenfest model and study Markov chains associated with it, and then look at transition matrix pseudospectra (see Exercises 15.1, 15.2).

The Ehrenfest model is a simple model for diffusion. Suppose that we have two urns, Urn 1 and Urn 2, with n balls split between them. At each time step we will choose a ball uniformly at random, move it to the other urn with probability $n/(n+1)$, or else leave it where it is. Denote by X_k the number of balls in Urn 1 at time k.

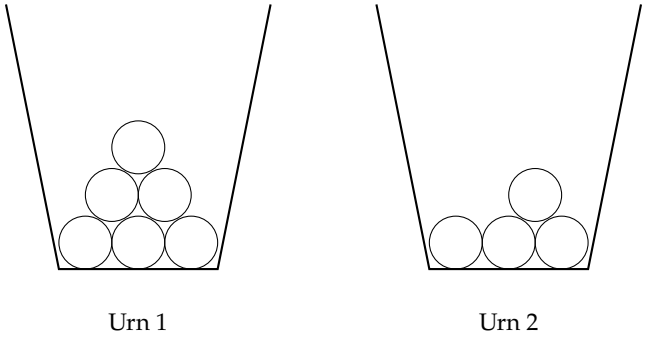


Figure 15.1 The Ehrenfest model with two urns and 10 balls.

(a) Explain briefly why the transition matrix P for $\{X_k\}$ is given by

$$P = \begin{pmatrix} \frac{1}{n+1} & \frac{n}{n+1} & & & & & \\ \frac{1}{n+1} & \frac{1}{n+1} & \frac{n-1}{n+1} & & & & \\ & \frac{2}{n+1} & \frac{1}{n+1} & \frac{n-2}{n+1} & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \frac{n-1}{n+1} & \frac{1}{n+1} & \frac{1}{n+1} & \\ & & & & \frac{n}{n+1} & \frac{1}{n+1} & \end{pmatrix}.$$

(b) Show that $\{X_k\}$ is irreducible and aperiodic, and so the conclusion of Theorem 15.17 holds.

(c) Show that the invariant distribution μ of $\{X_k\}$ is given by

$$\mu_j = 2^{-n} \binom{n}{j}, \quad j = 0, \ldots, n.$$

(d) Verify that P is not normal. Implement P in a programming language of your choice and compute (using `EigTool`, `PseudoPy` or otherwise) a number of ε-pseudospectra of P in the cases $n = 10, 100, 1000$. Similarly to Exercise 15.1(d), plot the contours of these pseudospectra for a variety of levels $\varepsilon \in (0, 1)$.

(e) For each $n = 10, 100, 1000$, compute and plot $\|P^m\|_p$ versus m for $p \in \{1, 2, \infty\}$ and a large range of m (until converged). Do the same for $\|A^m\|_p$, where $A = P - P^\infty$. How do the choice of norm and the dimension of S affect the behavior in m?

(f) Implement the Markov chain $\{X_k\}$ with starting point $X_0 = \lfloor n/2 \rfloor$. Fix $n = 10$, and choose a (non-constant) function $F\colon \{0, 1, \ldots, n\} \to \mathbb{R}$. Let $f \in \mathbb{R}^{n+1}$ be the corresponding vector $f_j = F(j)$ for each j. Given $N \in \mathbb{N}$, define $\hat{Z}_N$ by

$$\hat{Z}_N = \left| \langle f, \mu \rangle_2 - \frac{1}{N} \sum_{k=1}^{N} F(X_k) \right|.$$

Plot $\hat{Z}_N$ versus N. By plotting on logarithmic axes or otherwise, estimate the rate of convergence of $\hat{Z}_N$. Compare with the rate of convergence of

$$Z_N = \left| \langle f, \mu \rangle_2 - \frac{1}{N} \sum_{k=1}^{N} F(U_k) \right|,$$

where $\{U_k\}$ are independent samples from μ.
Note: Samples from μ can be generated using the `binornd` or `numpy.random.binomial`.
In the case F is defined by $F(j) = j$, compare both $\hat{Z}_N, Z_N$ with the theoretical 95% confidence bound for Z_N:

$$Z_N \leq \frac{1.96\sqrt{n}}{2\sqrt{N}}.$$

15.4 *For this exercise, use the dataset from Pseudospectra supplement. You can use either* `data.mat` *for MATLAB code,* `data.npy` *for Python code or* `data.csv` *for any other language; they contain the same dataset.*

This exercise carries on with Markov chains and pseudospectra (see Exercises 15.1, 15.2). Here we consider the *state-dependent random walk* on the set $S = \{0, \ldots, n\}$.

Let $p_0, \ldots, p_n \in (0, 1)$ be a sequence of probabilities. Define the Markov chain $\{X_k\}$ by

$$X_{k+1} = \begin{cases} X_k + 1 & \text{with probability } p_{X_k} \\ X_k - 1 & \text{with probability } 1 - p_{X_k} \end{cases}$$

where arithmetic is performed modulo $n + 1$.

(a) Write down the transition matrix P for $\{X_k\}$.
(b) Verify that P is not normal if the probabilities p_j are distinct. Implement P, generating the sequence of probabilities $p_0, \ldots, p_n \sim$

$U(0,1)$ i.i.d. Using one of the computational packages (Exercise 15.1), compute and plot a number of ε-pseudospectra of P in the case $n = 100$.

(c) We will now compare the properties of the chain when S has odd or even cardinality, specifically when $n = 50$ or $n = 51$ respectively. Load `data.npy`, `data.mat` or `data.csv`. The variable `p` contains $p_0, \ldots, p_{51} \sim U(0,1)$ i.i.d. Construct $P \in \mathbb{R}^{(n+1)\times(n+1)}$ for $n = 50$ using $p_0, \ldots, p_{50}$ and for $n = 51$ using $p_0, \ldots, p_{51}$.

(i) For $n = 50, 51$, compute and plot $\|P^m\|_p$ versus m for $p \in \{1, 2, \infty\}$ and m up to 10^6. Compare the behavior of the the norms for the two cases.

(ii) For $n = 50, 51$, using one of the packages, compute a number of ε-pseudospectra of $(P^{10^6})^*$. How do the pseudospectra compare between the two cases?

(iii) Plot the columns of $(P^{10^6})^*$ on the same axes. What difference do you notice between the cases $n = 50$ and $n = 51$?

(iv) For each $n = 50, 51$, take a column μ from $(P^{10^6})^*$. Is it the case that $P^*\mu = \mu$?

(v) Do you think that the chain $\{X_k\}$ is ergodic in either of the cases $n = 50, 51$? Theorize why this is the case, making reference to irreducibility and aperiodicity.

15.5 Prove Theorem 15.4, by expanding arbitrary vector $x \in \mathbb{C}^n$ in an appropriately chosen orthonormal basis, similar to part of the proof of Theorem 15.3.

15.6 State and prove a theorem substantiating Remark 15.12.

15.7 Compute all eigenvalues and eigenvectors of the matrix

$$A \coloneqq \begin{pmatrix} \lambda & \varepsilon \\ 0 & \lambda \end{pmatrix}.$$

for $\varepsilon = 0$ and $\varepsilon > 0$. Find the JNF for the matrix in both of these cases. Discuss your findings.

16
Jordan Normal Form and Applications

In this lecture we start by defining the action of function f on an arbitrary Jordan block, recalling from Lecture 15 that this enables us to define the action of function f on arbitrary matrix A, through its Jordan normal form. Using this, we then derive various properties of matrix functions and in particular of norms of A^k and e^{At}.

16.1 Functions of Jordan Blocks

Let $A \in \mathbb{C}^{n\times n}$. Recall that $A = SJS^{-1}$ for some $S \in \mathbb{C}^{n\times n}$ invertible and

$$J = \text{blockdiag}\{J_{n_1}(\lambda_1), \ldots, J_{n_k}(\lambda_k)\},$$

where

$$J_k(\lambda) = \begin{pmatrix} \lambda & 1 & & \mathbf{0} \\ & \ddots & \ddots & \\ & & & 1 \\ \mathbf{0} & & & \lambda \end{pmatrix} \in \mathbb{C}^{k\times k}$$

and $\sum_{j=1}^{k} n_j = n$. Then $f(A) = Sf(J)S^{-1}$, where

$$f(J) = \text{blockdiag}\{f(J_{n_1}(\lambda_1)), \ldots, f(J_{n_k}(\lambda_k))\}.$$

Remark 16.1. *Recall from Remark 15.14 that:*

(i) $\underbrace{f(B) = \sum_j a_j B^j}_{f:\mathbb{C}^{n\times n}\to\mathbb{C}^{n\times n}}$ *if* $\underbrace{f(z) = \sum_j a_j z^j}_{f:\mathbb{C}\to\mathbb{C}}$;

(ii) *to define $f(B)$, we need only define $f(J_k(\lambda))$.*

We also note that a Jordan block $J_k(\lambda) \in \mathbb{C}^{k\times k}$ may be written as $J_k(\lambda) = \lambda I + E$. Here $E \in \mathbb{C}^{k\times k}$ is defined by its action on $A \in \mathbb{C}^{k\times k}$ with rows $\{a_j\}$ as follows:

$$A = \begin{pmatrix} a_1^\top \\ a_2^\top \\ \vdots \\ a_{k-1}^\top \\ a_k^\top \end{pmatrix}, \quad EA = \begin{pmatrix} a_2^\top \\ a_3^\top \\ \vdots \\ a_k^\top \\ 0 \end{pmatrix}. \tag{16.1}$$

Theorem 16.2. *Assume that $f \in C^{k-1}(U;\mathbb{C})$ for U a neighborhood of the complex origin 0. The application of matrix function f to a Jordan block $f(J_k(\lambda))$ is an upper-triangular matrix with the form*

$$f(J_k(\lambda)) = \begin{pmatrix} f(\lambda) & f'(\lambda) & \frac{1}{2!}f''(\lambda) & \cdots & \frac{1}{(k-1)!}f^{(k-1)}(\lambda) \\ & \ddots & \ddots & \ddots & \vdots \\ & & & & \frac{1}{2!}f''(\lambda) \\ & & & & f'(\lambda) \\ & & & & f(\lambda) \end{pmatrix}. \tag{16.2}$$

Proof We will only prove the result in the case $k = 2$. However, the perturbation argument we use may be generalized to any $k \geq 2$. Let

$$J_2(\lambda, \epsilon) \coloneqq \begin{pmatrix} \lambda & 1 \\ 0 & \lambda + \epsilon \end{pmatrix}$$

so that $J_2(\lambda, 0) = J_2(\lambda)$. Then note

$$J_2(\lambda, \epsilon) = \begin{pmatrix} 1 & 1 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} \lambda & 0 \\ 0 & \lambda + \epsilon \end{pmatrix} \begin{pmatrix} 1 & -\epsilon^{-1} \\ 0 & \epsilon^{-1} \end{pmatrix}.$$

This is a diagonalization of $J_2(\lambda, \epsilon)$ with eigenvector-eigenvalue pairs

$$\left\{ \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \lambda \right\}, \quad \left\{ \begin{pmatrix} 1 \\ \epsilon \end{pmatrix}, \lambda + \epsilon \right\}.$$

Now note that

$$\begin{aligned} f(J_2(\lambda, \epsilon)) &= \begin{pmatrix} 1 & 1 \\ 0 & \epsilon \end{pmatrix} \begin{pmatrix} f(\lambda) & 0 \\ 0 & f(\lambda + \epsilon) \end{pmatrix} \begin{pmatrix} 1 & -\epsilon^{-1} \\ 0 & \epsilon^{-1} \end{pmatrix} \\ &= \begin{pmatrix} f(\lambda) & \epsilon^{-1}\left(f(\lambda + \epsilon) - f(\lambda)\right) \\ 0 & f(\lambda + \epsilon) \end{pmatrix} \end{aligned}$$

Hence

$$f(J_2(\lambda)) = \lim_{\epsilon \to 0} f(J_2(\lambda, \epsilon)) = \begin{pmatrix} f(\lambda) & f'(\lambda) \\ 0 & f(\lambda) \end{pmatrix}.$$

as desired. □

16.2 A^k and e^{At}

Theorem 16.3. *Let $A \in \mathbb{C}^{n \times n}$. Assume that $\rho(A) < 1$. Then there is $\alpha \in (0, 1)$ and induced matrix norm $\| \cdot \|$ such that $\|A^k\| \leq \alpha^k$.*

Proof Let $\rho(A) = \beta$ and chose $\delta = (1 - \beta)/2$. Then, by Theorem 15.11, there is a norm $\| \cdot \|$ in which

$$\|A\| \leq \rho(A) + \delta = \beta + \frac{1 - \beta}{2} = \frac{1 + \beta}{2} =: \alpha \in (0, 1).$$

Because $\| \cdot \|$ is an induced matrix norm, we deduce from Theorem 15.2 that

$$\|A^k\| \leq \|A\|^k \leq \alpha^k.$$

□

Theorem 16.4. *The matrix differential equation*

$$\dot{x} = Ax, \quad x(0) = a$$

has solution $x(t) = e^{tA}a$.

Proof Substituting the proposed solution into the differential equation and noting that a is arbitrary we see that it suffices to show that

$$\frac{d}{dt}\{e^{tA}\} = Ae^{tA}.$$

By Remark 16.1 it suffices to show that

$$\frac{d}{dt}\{e^{tJ_k(\lambda)}\} = J_k(\lambda)e^{tJ_k(\lambda)}.$$

From (16.2) we see that $e^{tJ_k(\lambda)}$ is an upper-triangular matrix with the

form

$$e^{tJ_k(\lambda)} = \begin{pmatrix} e^{t\lambda} & te^{t\lambda} & \frac{1}{2!}t^2e^{t\lambda} & \cdots & \frac{1}{(k-1)!}t^{(k-1)}e^{t\lambda} \\ & \ddots & \ddots & \ddots & \vdots \\ & & & & \frac{1}{2!}t^2e^{t\lambda} \\ & & & & te^{t\lambda} \\ & & & & e^{t\lambda} \end{pmatrix}.$$

Each matrix entry has the form

$$\frac{1}{\ell!}t^\ell e^{t\lambda}$$

and the derivative with respect to t of this function is

$$\lambda\frac{1}{\ell!}t^\ell e^{t\lambda} + \frac{1}{(\ell-1)!}t^{\ell-1}e^{t\lambda}.$$

From this it follows that

$$\frac{d}{dt}\{e^{tJ_k(\lambda)}\} = \lambda e^{tJ_k(\lambda)} + Ee^{tJ_k(\lambda)}$$

where E is defined in (16.1). But $\lambda I + E = J_k(\lambda)$ and the result is proved. □

Theorem 16.5. *Let $A \in \mathbb{C}^{n\times n}$. Assume that there is $\beta \in (0,\infty)$ such that $Re(\lambda_j) < -\beta$ for all eigenvalues λ_j of A. Then there is a constant $C > 0$ such that*

$$\|e^{At}\|_\infty \le Ce^{-\beta t}.$$

Proof Recall that the JNF shows that

$$e^{At} = Se^{Jt}S^{-1}, \quad \|e^{At}\|_\infty \le \|S\|_\infty\|S^{-1}\|_\infty\|e^{Jt}\|_\infty.$$

Thus it suffices to prove the result for e^{Jt}. Now

$$e^{Jt} = \text{blockdiag}\{\dots, e^{J_{n_j}(\lambda_j)t}, \dots\}.$$

Thus it suffices to find

$$\max_{1\le j\le k} \|e^{J_{n_j}(\lambda_j)t}\|_\infty.$$

Let $f(z) = e^{tz}$. Now $e^{J_{n_j}(\lambda_j)t}$ has row sum bounded by

$$\sum_{\ell=0}^{n_j-1} \left|\frac{1}{\ell!}f^{(\ell)}(\lambda_j)\right|,$$

and note that $f^{(\ell)}(\lambda) = t^\ell e^{\lambda t}$. Thus the row sum is bounded by

$$n_j(\max\{1, |t|\})^{n_j-1}|e^{\lambda_j t}|.$$

But $|e^{\lambda_j t}| \le e^{-\beta t}$ by supposition, so there exists a constant $C_j' > 1$ such that

$$\sup_{|t|>1} |t|^{n_j-1}|e^{\lambda_j t}|e^{\beta t} \le C_j'.$$

Hence, because this holds for each Jordan block and $n_j \le n$,

$$\|e^{Jt}\|_\infty \le \left(n \max_{1\le j\le k} C_j'\right)e^{-\beta t}.$$

□

16.3 Cayley–Hamilton Theorem

Definition 16.6. *The* characteristic polynomial *of a matrix $A \in \mathbb{C}^{n\times n}$ is*

$$\chi_A(z) = \det(A - zI).$$

Proposition 16.7. *The eigenvalues of A are the zeros of $\chi_A(z)$.*

Definition 16.8. *If $\lambda \in \mathbb{C}$ is an eigenvalue of A, then λ has* algebraic multiplicity *q if q is the largest integer such that $(z-\lambda)^q$ divides $\chi_A(z)$. The* geometric multiplicity *r is the dimension of* $\mathrm{Ker}(A - \lambda I)$.

Remark 16.9. *The following properties of the JNF are needed in what follows.*

(i) *If $q = r \ge 1$ for eigenvalue λ, then there are q 1x1 Jordan blocks with diagonal λ.*

(ii) *If $q > r = 1$ for eigenvalue λ, then there is 1 qxq Jordan block with diagonal λ.*

(iii) *If $q > r \ge 1$ for eigenvalue λ, then the resulting Jordan blocks with diagonal entry λ have dimensions summing to q.*

(iv)
$$\frac{d^l}{dz^l}(\chi_A(z))\bigg|_{z=\lambda_j} = 0, \quad l = 0, \dots, n_j - 1. \tag{16.3}$$

This follows from the fact that

$$\frac{d^l}{dz^l}(\chi_A(z))\bigg|_{z=\lambda_j} = 0, \quad l = 0, \dots, q - 1$$

and first three comments show that $n_j \le q$.

Theorem 16.10 (Cayley-Hamilton). $\chi_A(A) = 0$.

Proof It suffices to show that

$$\chi_A(J_{n_j}(\lambda_j)) = 0, \quad j = 1, \ldots, k.$$

This holds by the preceding theorem and (16.3). □

Exercises

16.1 Let J denote the standard 2×2 Jordan block with diagonal entry λ. Consider the iteration

$$x_{k+1} = Jx_k + u$$

for some fixed vectors u and x_0. If $|\lambda| < 1$, prove that x_k converges as $k \to \infty$ and identify the limit. What can you say about the limiting behavior of x_k as $k \to \infty$ if $\lambda = 1$ and if $\lambda = -1$?

16.2 Let J denote the standard 3x3 Jordan block with diagonal entry λ. Compute $\exp(Jt)$, $\sin(Jt)$ and $\cos(Jt)$. Using these formulae, solve the following differential equations:

(a) $\dot{u} = Ju$, $u(0) = u_0$;
(b) $\ddot{u} = -J^2u$, $u(0) = u_0$, $\dot{u}(0) = v_0$.

16.3 Let $A \in \mathbb{C}^{n\times n}$ be a matrix for which all eigenvalues have real part negative. Is it true that $\exp(At)$ is bounded independently of $t > 0$? If so prove it; if not, find a counter example.

16.4 Let $A \in \mathbb{C}^{n\times n}$. Using the Jordan normal form, show that $A^k \to 0$ as $k \to \infty$ if and only if $\rho(A) < 1$. Furthermore, show that $\|A^k\|$ is unbounded as $k \to \infty$ if $\rho(A) > 1$.

16.5 Using Exercise 16.4, prove Gelfand's formula which states that, for every $A \in \mathbb{C}^{n\times n}$,

$$\rho(A) = \lim_{k\to\infty} \|A^k\|^{1/k}.$$

16.6 For each $\varepsilon \geq 0$, define the matrix $A_\varepsilon \in \mathbb{R}^{2\times 2}$ by

$$A_\varepsilon = \begin{pmatrix} \varepsilon & 0 \\ 1 & 0 \end{pmatrix}.$$

Calculate the Jordan normal form J_ε of A_ε for each $\varepsilon \geq 0$. Deduce that $A_\varepsilon \to A_0$, but $J_\varepsilon \not\to J_0$ as $\varepsilon \to 0$ for any arrangement of Jordan blocks in J_ε.

16.7(a) Let $A, B \in \mathcal{L}(X, X)$ be bounded operators on a Banach space X such that $AB = BA$. Show that

$$e^{A+B} = e^A e^B,$$

and hence e^A is always invertible.

(b) Let $A, B \in \mathcal{L}(X, X)$ be arbitrary bounded operators on a Banach space X.

(i) Show that when $X = \mathbb{R}$, it is always the case that $AB = BA$ and so $e^{A+B} = e^A e^B$.

(ii) By considering the case $X = \mathbb{R}^2$, show that in general

$$e^{A+B} \neq e^A e^B.$$

(iii) Show that

$$e^{A+B} = \lim_{n\to\infty} (e^{A/n} e^{B/n})^n$$

where convergence is in the operator norm.
Hint: you may wish to consider the operators $T_n = e^{(A+B)/n}$ and $S_n = e^{A/n} e^{B/n}$, and use the telescoping sum

$$T_n^n - S_n^n = T_n^n - S_n T_n^{n-1} + S_n T_n^{n-1} - S_n^2 T_n^{n-2} + S_n^2 T_n^{n-2} - \ldots - S_n^n.$$

17
Integral Operators and Applications

In this lecture, we apply several core topics in linear analysis, including Sobolev spaces, compactness, and spectral theory, in the context of the one-dimensional linear Poisson equation. In particular, we derive and study the properties of its solution operator, a compact and symmetric integral operator, through the use of Green's functions. We then generalize the ideas developed thus far to formulate an abstract version of a *nonlinear* boundary value problem; this example will then be used in Lecture 18.

17.1 Motivating Problem

We consider the following *Poisson problem*, which is a special case of the linear ordinary differential equation (BVP) from Lecture 12 with $\kappa(x) \equiv 1$ for all $x \in I := (0,1)$:

$$\begin{aligned} -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x) &= r(x), \quad x \in I \\ u(x) &= 0, \quad x \in \partial I := \{0,1\}\,. \end{aligned} \tag{PBVP}$$

Note that, as presented, (PBVP) is defined to hold pointwise on the spatial domain I. Instead, we can formulate an abstract version of the boundary value problem that turns it into a linear operator equation on the Hilbert space $X := (L^2(I;\mathbb{R}), \langle\cdot,\cdot\rangle_X, \|\cdot\|_X)$, where $\langle\cdot,\cdot\rangle_X$, $\|\cdot\|_X$ denote the usual L^2 inner product and norm.

Indeed, define the operator (interpreted in terms of weak derivatives)

$$\begin{aligned}\mathcal{A} : D(\mathcal{A}) &\to X \\ v &\mapsto \mathcal{A}v := -\frac{\mathrm{d}^2 v}{\mathrm{d}x^2},\end{aligned}$$

where

$$D(\mathcal{A}) := H^2(I;\mathbb{R}) \cap H_0^1(I;\mathbb{R}) \subset X$$

encodes the zero Dirichlet boundary conditions and is a dense subset of X. The abstract form of (PBVP) is then as follows: given $r \in X$, find $u \in D(\mathcal{A})$ so that

$$\mathcal{A}u = r. \tag{PBVP-a}$$

Remark 17.1. *$\mathcal{A}$, as a differential operator, is an example of a* densely-defined unbounded *linear operator; the choice of its domain $D(\mathcal{A})$ is crucial for $\mathcal{A}$ to be fully specified: it contains the information about the boundary conditions to be satisfied by solutions of* (PBVP)*; and it contains information about the number of (weak) derivatives required of the solution – two in this case.*

Definition 17.2. *The* Green's function *associated to $\mathcal{A}$ is $g : I \times I \to \mathbb{R}$,*

$$g(x;y) = \begin{cases}(1-y)x, & 0 < x < y < 1 \\ y(1-x), & 0 < y < x < 1.\end{cases}$$

Remark 17.3. *Note that g is symmetric, i.e., $g(x;y) = g(y;x)$ for all $x, y \in I$, and a more concise way to express g is via $g(x;y) = \min\{x,y\} - xy$. It may be found by solving the equation*

$$\mathcal{A}g(\cdot, y) = \delta_y,$$

where $\delta_y = \delta(\cdot - y)$ is the Dirac delta distribution centered at $y \in I$. The solution concept with right hand side a distribution needs careful consideration. From an operational point of view the equation is solved by solving $\mathcal{A}g(\cdot, y) = 0$ on either side of the point $x = y$ and imposing that the solution $g(\cdot, y)$ is continuous at y and that the derivative $\frac{d}{dx}g(\cdot, y)$ jumps downwards by 1 unit across $x = y$.

Definition 17.4. *The integral operator $G : X \to X$ associated to Green's function g is given by*

$$(Gv)(x) = \int_0^1 g(x;y)v(y)\mathrm{d}y$$

for every $v \in X$.

Remark 17.5. *The Green's function is also known as the* kernel *of the integral operator G.*

17.2 Properties of Integral Operator

In this section, we will prove the following facts about G:

Remark 17.6.
- $G \in \mathcal{L}(X, X)$;
- $\mathcal{A}G = I$ *on* X *and* $G\mathcal{A} = I$ *on* $D(A)$*;*
- $u = Gr$ *solves* (PBVP-a)*;*
- G *is symmetric;*
- $G \in \mathcal{L}(X, X)$ *is compact.*

Proposition 17.7. *It holds that* $G \in \mathcal{L}(X, X)$ *and* $\|G\|_{\mathcal{L}(X,X)} \leq 1$.

Proof We leave it to the reader to show linearity. For boundedness, notice that

$$\sup_{(x,y)\in I\times I} |g(x;y)| \leq 1.$$

Then by Cauchy-Schwarz,

$$\|Gv\|_X^2 = \int_0^1 \left(\int_0^1 g(x;y)v(y)\mathrm{d}y \right)^2 \mathrm{d}x \leq \int_0^1 \left(\int_0^1 v(y)^2 \mathrm{d}y \right) \mathrm{d}x = \|v\|_X^2.$$

Thus

$$\|G\|_{\mathcal{L}(X,X)} = \sup_{r\neq 0} \frac{\|Gv\|_X}{\|v\|_X} \leq 1$$

and we are done. □

Proposition 17.8. *It holds that* $\mathcal{A}G = I$*, that is* $\mathcal{A}Gr = r$ *for all* $r \in X$*. Thus* $u = Gr$ *solves* (PBVP-a)*.*

Proof We sketch the proof in the setting where r and the second derivative of u are continuous on $\bar{I}$; care is needed to extend the desired result for all $r \in X$ and $u \in D(A)$ by a density argument.

For r continuous on $\bar{I}$ we note that it is pointwise well-defined. We define $u = Gr$ and show that $(\mathcal{A}u)(x) = r(x)$ for all $x \in I$. Note that by construction $u(0) = u(1) = 0$ and so it remains to show that the second

derivative of u is $-r$ at each point in I. Writing out $u(x)$ explicitly, we have

$$u(x) = \int_0^x y(1-x)r(y)\mathrm{d}y + \int_x^1 (1-y)xr(y)\mathrm{d}y.$$

Then, differentiating once with the Leibniz rule,

$$\begin{aligned}\frac{\mathrm{d}u}{\mathrm{d}x}(x) &= x(1-x)r(x) - \int_0^x yr(y)\mathrm{d}y - (1-x)xr(x) + \int_x^1 (1-y)r(y)\mathrm{d}y \\ &= -\int_0^x yr(y)\mathrm{d}y + \int_x^1 (1-y)r(y)\mathrm{d}y.\end{aligned}$$

Differentiating once more,

$$-(\mathcal{A}u)(x) = \frac{\mathrm{d}^2u}{\mathrm{d}x^2}(x) = -xr(x) - (1-x)r(x) = -r(x).$$

□

Proposition 17.9. *It holds that $G\mathcal{A} = I$, that is $G\mathcal{A}u = u$ for all $u \in D(A)$.*

Proof We prove the result for twice continuously differentiable u in $\bar{I}$, satisfying $u(x) = 0$ on ∂I. The result may be extended to $D(A)$ by density. Using integration by parts twice, noting that the boundary contributions of the two integrals cancel in the first integration by parts,

$$\begin{aligned}(G\mathcal{A}u)(x) &= -\int_0^1 g(x;y)u''(y)dy \\ &= -\int_0^x y(1-x)u''(y)dy - \int_x^1 (1-y)xu''(y)dy \\ &= \int_0^x (1-x)u'(y)dy - \int_x^1 xu'(y)dy \\ &= (1-x)u(y)\Big|_0^x - xu(y)\Big|_x^1 \\ &= (1-x)u(x) + xu(x) \\ &= u(x).\end{aligned}$$

□

Proposition 17.10. *$G \in \mathcal{L}(X, X)$ is symmetric.*

Proof For any $r, q \in X$,

$$\begin{aligned}
\langle Gr, q\rangle_X &= \int_0^1 \left(\int_0^1 g(x;y) r(y) \mathrm{d}y \right) q(x) \mathrm{d}x \\
&= \int_0^1 \int_0^1 g(x;y) r(y) q(x) \mathrm{d}y \mathrm{d}x \\
&= \int_0^1 \int_0^1 g(y;x) r(x) q(y) \mathrm{d}x \mathrm{d}y \\
&= \int_0^1 \int_0^1 g(x;y) r(x) q(y) \mathrm{d}y \mathrm{d}x \\
&= \int_0^1 r(x) \left(\int_0^1 g(x;y) q(y) \mathrm{d}y \right) \mathrm{d}x \\
&= \langle r, Gq\rangle_X,
\end{aligned}$$

where we have used the symmetry of g and the Fubini-Tonelli theorem. □

Now consider the well-known Fourier sine basis $\{\varphi^{(j)}\}_{j\in\mathbb{N}}$ for X, defined by

$$\varphi^{(k)}(x) = \sqrt{2}\sin(k\pi x).$$

It is a fact (which can be proven with Fourier analysis) that $\{\varphi^{(j)}\}_{j\in\mathbb{N}}$ form an orthonormal basis for X. Since these functions are smooth and satisfy the boundary conditions, the action of $\mathcal{A}$ on the basis is

$$\mathcal{A}\varphi^{(j)} = (j\pi)^2 \varphi^{(j)}$$

for all $j \in \mathbb{N}$, which can be directly verified. Hence

$$\mathcal{A}((j\pi)^{-2}\varphi^{(j)}) = \varphi^{(j)}$$

and so by Proposition 17.9 we conclude that

$$G\varphi^{(j)} = (j\pi)^{-2}\varphi^{(j)}.$$

That is, the eigenvalue/eigenvector problem

$$\begin{aligned}
G\varphi &= \lambda\varphi, \\
\|\varphi\|_X^2 &= 1.
\end{aligned}$$

has solutions $\{(\lambda_j, \varphi^{(j)})\}_{j\in\mathbb{N}}$, where $\lambda_j = (j\pi)^{-2}$. Using this fact, we have the following proposition, which will be useful in Lecture 18.

Proposition 17.11.

$$\|G\|_{\mathcal{L}(X,X)} = \frac{1}{\pi^2}.$$

Proof Since the eigenvectors $\{\varphi^{(j)}\}_{j\in\mathbb{N}}$ form an orthonormal basis for X and G is symmetric,

$$\|G\|_{\mathcal{L}(X,X)} = \rho(G) = \lambda_1 = \frac{1}{\pi^2}.$$

□

Proposition 17.12. *$G \in \mathcal{L}(X, X)$ is compact.*

Proof We show that G maps X into a space that is compactly embedded in X. If $v \in X$, then

$$v = \sum_{j=1}^{\infty} \hat{v}_j \varphi^{(j)}, \quad \hat{v}_j = \langle \varphi^{(j)}, v \rangle_X,$$

and $\|v\|_X^2 = \|\hat{v}\|_{\ell^2(\mathbb{N};\mathbb{R})}^2$. With a slight abuse of notation, we introduce the norm $\|\hat{v}\|_{\mathcal{X}^s}^2 = \|v\|_{\mathcal{X}^s}^2 := \sum_{j=1}^{\infty} j^{2s}\hat{v}_j^2$, which allows us to write

$$\mathcal{X}^s = \{v \in X : \|v\|_{\mathcal{X}^s} < \infty\}.$$

From Lecture 10, we know that $\mathcal{X}^s$ is compactly embedded in X if $s > 0$. It suffices to show that $G \in \mathcal{L}(X, \mathcal{X}^t)$ for some $t > 0$.

Since we can write

$$Gv = \sum_{j=1}^{\infty} \frac{\hat{v}_j}{j^2\pi^2} \varphi^{(j)},$$

the norm in $\mathcal{X}^t$ is

$$\|Gv\|_{\mathcal{X}^t}^2 = \sum_{j=1}^{\infty} j^{2t} \frac{\hat{v}_j^2}{j^4\pi^4} = \frac{1}{\pi^4} \sum_{j=1}^{\infty} j^{2t-4}\hat{v}_j^2 \le \frac{1}{\pi^4} \sum_{j=1}^{\infty} \hat{v}_j^2 = \frac{1}{\pi^4}\|v\|_X^2$$

for $0 < t < 2$. Thus, $G \in \mathcal{L}(X, \mathcal{X}^t)$ for $0 < t < 2$ and hence $G \in \mathcal{L}(X, X)$ is a compact operator. □

Remark 17.13. *There are many other ways to prove compactness of G, including using Hilbert-Schmidt theory (Exercise 13.2) or directly showing that $G : X \to D(\mathcal{A})$, since $D(\mathcal{A}) \subset H_0^1 \hookrightarrow L^2$ compactly embeds into L^2. We have shown this rather direct computation so that we may link directly to Lecture 10.*

Example 17.14. Since we have shown $G \in \mathcal{L}(X, X)$ is a compact symmetric operator, we may apply the spectral theory in Section 13.3 to expand

$$G = \sum_{j=1}^{\infty} (j\pi)^{-2} \varphi^{(j)} \otimes \varphi^{(j)}.$$

Further, by Section 13.4, for any $\ell \in \mathbb{N}$ a finite-rank approximation to G is

$$G_\ell = \sum_{j=1}^{\ell} (j\pi)^{-2} \varphi^{(j)} \otimes \varphi^{(j)}.$$

It also follows that $\{G_\ell\}_{\ell \in \mathbb{N}}$ is Cauchy in $\mathcal{L}(X, X)$. To see this, suppose $m \geq n \in \mathbb{N}$ without loss of generality. Then,

$$\begin{aligned} \|G_m - G_n\|_{\mathcal{L}(X,X)} &\leq \sum_{j=n+1}^{m} (j\pi)^{-2} \|\varphi^{(j)} \otimes \varphi^{(j)}\|_{\mathcal{L}(X,X)} \\ &\leq \sum_{j=n+1}^{\infty} (j\pi)^{-2} \to 0 \quad \text{as } n \to \infty \end{aligned}$$

since the series is convergent. ◊

17.3 Nonlinear Problem

For a given function $f : \mathbb{R} \to \mathbb{R}$, consider the following problem:

$$\begin{aligned} -\frac{\mathrm{d}^2 u}{\mathrm{d}x^2}(x) &= f(u(x)), \quad x \in I \\ u(x) &= 0, \quad x \in \partial I. \end{aligned} \tag{NPBVP}$$

Defining Nemytskii operator $F : X \to X$ by $(F(u))(x) \coloneqq f(u(x))$, we may rewrite (NPBVP) in the following form:

$$\mathcal{A}u = F(u) \tag{NPBVP-a}$$

Using Proposition 17.9 we deduce that $u = G\mathcal{A}u = GF(u)$ for all $u \in D(\mathcal{A})$. Thus solving (NPBVP-a) is equivalent to solving the nonlinear fixed point equation

$$u = T(u), \tag{FP}$$

where $T : X \to X$ is given by $T \coloneqq G \circ F$. Thus we have reformulated the original problem (NPBVP) as an abstract fixed point problem on Hilbert

space X. In the next lecture, Lecture 18, we will study the existence and uniqueness of solutions to (FP) using the contraction mapping principle.

Exercises

17.1 Let $A \in \mathbb{R}^{n\times n}$ and let $e^{(j)} \in \mathbb{R}^n$ denote the vector with entries $e_i^{(j)} = \delta_{ij}$. Let the *discrete Green's functions* $g^{(j)}$ solve $Ag^{(j)} = e^{(j)}$. Express the solution of the equation $Au = r$ in terms of the discrete Green's functions.

18
Contraction Mapping Theorem

In this lecture we make the first of several forays into *nonlinear* analysis, studying the contraction mapping theorem and its application to ordinary differential equations (ODEs). We conclude the lecture by studying the uniform contraction mapping principle; in words, this demonstrates that certain properties of the contraction map, such as continuity with respect to a parameter, are inherited by the fixed point. Throughout this lecture let $(X, \|\cdot\|)$ be a Banach space over $\mathbb{R}$.

18.1 Lipschitz Functions

We define two notions of Lipschitz in X. We let $B(v, r)$ denote the ball of radius r, centered at v, in X. Thus $B(v, r) = B_X(v, r)$, a useful economy of notation.

Definition 18.1. $f : X \to X$ *is* Lipschitz *if* $\exists L \in (0, \infty)$ *such that*

$$\sup_{\substack{u,v \in X \\ u \neq v}} \frac{\|f(u) - f(v)\|}{\|u - v\|} \leq L.$$

$f : X \to X$ *is* locally Lipschitz *if, for every* $R > 0$, $\exists L = L(R) \in (0, \infty)$ *such that*

$$\sup_{\substack{u,v \in B(0,R) \\ u \neq v}} \frac{\|f(u) - f(v)\|}{\|u - v\|} \leq L(R).$$

Remark 18.2. *We say L is the Lipschitz constant. We write Lipschitz(L) for Lipschitz (or locally Lipschitz) with Lipschitz constant L.*

Lemma 18.3. *Let $F : X \to X$ be Lipschitz(L) and let $G \in \mathcal{L}(X, X)$. Then $G \circ F$ is Lipschitz$(\|G\|_{\mathcal{L}(X,X)} L)$.*

The proof is left as Exercise 18.3.

18.2 Main Theorem and Proof

Let $T : X \to X$ be a (possibly nonlinear) mapping.

Definition 18.4. *Let M be a closed non-empty set in X. Then $T : X \to X$ is a* contraction on M *(or is* contractive on M*) if $\exists \lambda \in (0, 1)$ such that*

$$\sup_{\substack{u,v \in M \\ u \neq v}} \frac{\|T(u) - T(v)\|}{\|u - v\|} \leq \lambda.$$

Remark 18.5. *Note the following:*

- *The concept applies in the setting $M = X$ as a special case.*
- *If T is linear, then it is a contraction on X if and only if $\|T\|_{\mathcal{L}(X,X)} \leq \lambda \in (0, 1)$.*
- *If T is Lipschitz(L) on X and $L < 1$, then it is a contraction on X.*
- *If T is locally Lipschitz(L) on $B(v, R)$ and $L < 1$, then it is a contraction on $B(v, R)$.*

Theorem 18.6 (Contraction Mapping Principle)**.** *Assume that T maps M into itself and is a contraction on M with constant $\lambda \in (0, 1)$. Then the equation*

$$u = T(u)$$

has a solution $u \in M$ and the solution is unique within M. Furthermore, the iteration

$$u_{n+1} = T(u_n)$$

satisfies, if $u_0 \in M$,

$$\|u_n - u\| \leq \lambda^n (1 - \lambda)^{-1} \|u_1 - u_0\| \quad \forall n \in \mathbb{Z}^+.$$

Proof We first prove existence, uniqueness and convergence of the iteration, starting by showing that $\{u_n\}$ is Cauchy in X. To this end, note that

$$\begin{aligned} \|u_{n+1} - u_n\| &= \|T(u_n) - T(u_{n-1})\| \leq \lambda \|u_n - u_{n-1}\| \\ &= \lambda \|T(u_{n-1}) - T(u_{n-2})\| \leq \lambda^2 \|u_{n-1} - u_{n-2}\| \\ &\leq \cdots \leq \lambda^n \|u_1 - u_0\|. \end{aligned}$$

It follows that

$$\begin{aligned}\|u_n - u_{n+m}\| &= \|(u_n - u_{n+1}) + \cdots + (u_{n+m-1} - u_{n+m})\| \\ &\leq \|u_n - u_{n+1}\| + \cdots + \|u_{n+m-1} - u_{n+m}\| \\ &\leq (\lambda^n + \cdots + \lambda^{n+m-1})\|u_1 - u_0\| \\ &\leq \lambda^n(1 + \lambda + \lambda^2 + \cdots)\|u_1 - u_0\| \\ &= \lambda^n(1-\lambda)^{-1}\|u_1 - u_0\|.\end{aligned}$$

Thus

$$\|u_n - u_{n+m}\| \leq \lambda^n(1-\lambda)^{-1}\|u_1 - u_0\|. \tag{18.1}$$

Since $\lambda \in (0,1)$, we deduce that the sequence is Cauchy and hence has a limit $u \in X$. We now show this limit point is a solution of our equation. First note that, by induction, the entire sequence $\{u_n\}$ is in M. Since M is closed, the limit u is in M, from which it follows that $T(u) \in M$. As a consequence

$$\|T(u_n) - T(u)\| \leq \lambda\|u_n - u\|$$

and since the right hand side tends to zero as $n \to \infty$, it follows that $T(u_n) \to T(u)$ as $n \to \infty$. Now note that

$$\begin{aligned}\|u - T(u)\| &\leq \|u - u_{n+1}\| + \|u_{n+1} - T(u)\| \\ &= \|u - u_{n+1}\| + \|T(u_n) - T(u)\|.\end{aligned}$$

Thus letting $n \to \infty$ in the right hand side we deduce that $u = T(u)$ as required.

For the purpose of contradiction, assume that we have two fixed points u and v, both in M. Then

$$\|u - v\| = \|T(u) - T(v)\| \leq \lambda\|u - v\|$$

and since $\lambda \in (0,1)$, this implies that $\|u - v\| = 0$ and hence that $u = v$. To obtain the error estimate we let $m \to \infty$ in (18.1), giving the desired result. □

18.3 Application to ODE Boundary Value Problem

Here we consider the ordinary differential equation boundary value problem (NPBVP) from Lecture 17, reformulated as the fixed point equation (FP). We adopt the notation from that section and make the following assumption:

Assumption 18.7. *Function $f : \mathbb{R} \to \mathbb{R}$ is Lipschitz(L).*

Lemma 18.8. *Let Assumption 18.7 hold. The Nemytskii operator $F : X \to X$ and is Lipschitz(L).*

Proof First we note that the following property holds:

$$\begin{aligned}\|F(u) - F(v)\|_X^2 &= \int_0^1 \left|f\big(u(x)\big) - f\big(v(x)\big)\right|^2 dx \\ &\le \int_0^1 L^2 \left|u(x) - v(x)\right|^2 dx \\ &= L^2 \|u - v\|_X^2.\end{aligned}$$

This implies F is Lipschitz(L) if we can show that $F : X \to X$. Now let $f(0) = f_0$ and note that $(F(0))(x) = f_0$ for all $x \in I$. Using the fact that $\|F(0)\|_X = f_0$ and the above property, we deduce that

$$\begin{aligned}\|F(u)\|_X &\le \|F(u) - F(0)\|_X + \|F(0)\|_X \\ &\le L\|u\|_X + f_0,\end{aligned}$$

showing that F indeed maps X to itself. □

Theorem 18.9. *Consider the fixed point equation* (FP) *from Lecture 17 and let Assumption 18.7 hold. If $L < \pi^2$, then there is a unique solution $u \in X$ to* (FP).

Proof By Proposition 17.11 and Lemma 18.8, it follows, by application of Lemma 18.3, that $T = G \circ F$ is Lipschitz($\pi^{-2}L$) as a mapping from X into itself. The Contraction Mapping Principle proves existence of a unique solution in X if $\pi^{-2}L < 1$. □

18.4 Application to ODE Initial Value Problem

Now consider the ordinary differential equation

$$\frac{du}{dt}(t) = f(u(t)), \quad u(0) = u_0. \tag{ODE}$$

Any function u which is a solution of this equation is also a solution of the integral equation

$$u(t) = u_0 + \int_0^t f(u(s))\, ds. \tag{IE}$$

The converse is true if the function u solving (IE) is continuously differentiable. Thus, we study (IE) in what follows. The following provides the Banach space setting natural for this problem.

Definition 18.10. *Let $\|\cdot\|_{\mathbb{R}^m}$ be a norm on $\mathbb{R}^m$. Consider*

$$X := C([0,T];\mathbb{R}^m) = \{u : [0,T] \to \mathbb{R}^m \mid u \text{ is continuous}\}$$

and $\|\cdot\| = \|\cdot\|_X$ given by

$$\|u\| = \sup_{t\in[0,T]} \|u(t)\|_{\mathbb{R}^m}.$$

Theorem 18.11. *$(X, \|\cdot\|_X)$ as defined above is a Banach space. Choosing different norms on $\mathbb{R}^m$ results in an equivalent norm on X.*

In what follows, we view u_0 both as an element of $\mathbb{R}^m$ and as an element of X defined by

$$u_0(t) := u_0, \quad t \in [0,T].$$

Fix $r > 0$. Assume $f : \mathbb{R}^m \to \mathbb{R}^m$ is locally Lipschitz. Define

$$M = \overline{B_X(u_0, r)} = \{u \in X : \|u - u_0\| \le r\}.$$

Choose $R > r + \|u_0\|_{\mathbb{R}^m}$, so that $M \subseteq B_X(0,R)$. Let $L(R)$ be the Lipschitz constant for f on $B(0,R)$ and then choose $T^+ > 0$ sufficiently small so that

$$T^+ L(R) \le \frac{1}{2}, \quad T^+ \|f(u_0)\|_{\mathbb{R}^m} \le \frac{r}{2}.$$

Theorem 18.12. *Let f be locally Lipschitz. Define X with $T = T^+$. Then* (IE) *has a unique solution in M.*

Proof Define $F : X \to X$ by

$$(Fu)(t) = u_0 + \int_0^t f(u(s))\,ds, \quad 0 \le t \le T^+,$$

and note that a solution of (IE) is a fixed point of F. Note also that any locally Lipschitz function is continuous, and in particular f is continuous. Thus, since $s \mapsto u(s)$ is continuous, it follows that $s \mapsto f(u(s))$ is continuous. Furthermore, $s \mapsto f(u(s))$ continuous implies that $t \mapsto \int_0^t f(u(s))\,ds$ is continuous. Thus, we deduce that indeed $F : X \to X$.

We will now show that $F : M \to M$. Assume $u \in M$. Then,

$$\begin{aligned}
\|F(u) - u_0\| &= \sup_{t\in[0,T^+]} \left\| \int_0^t f(u(s))\, ds \right\|_{\mathbb{R}^m} \\
&\leq \sup_{t\in[0,T^+]} \int_0^t \|f(u(s))\|_{\mathbb{R}^m} ds \\
&= \int_0^{T^+} \|f(u(s))\|_{\mathbb{R}^m} ds \\
&\leq \int_0^{T^+} \left(\|f(u_0)\|_{\mathbb{R}^m} + \|f(u(s)) - f(u_0)\|_{\mathbb{R}^m} \right) ds \\
&\leq T^+ \|f(u_0)\|_{\mathbb{R}^m} + \int_0^{T^+} L(R)\|u(s) - u_0\|_{\mathbb{R}^m} ds \\
&\leq T^+ \|f(u_0)\|_{\mathbb{R}^m} + \|u - u_0\| \int_0^{T^+} L(R)\, ds \\
&\leq T^+ \|f(u_0)\|_{\mathbb{R}^m} + T^+ L(R) r \\
&\leq \frac{r}{2} + \frac{r}{2} \\
&= r
\end{aligned}$$

as desired.

Finally, we show that $F : M \to M$ is a contraction:

$$\begin{aligned}
\|F(u) - F(v)\| &= \sup_{t\in[0,T^+]} \left\| \int_0^t (f(u(s)) - f(v(s)))\, ds \right\|_{\mathbb{R}^m} \\
&\leq \int_0^{T^+} \|f(u(s)) - f(v(s))\|_{\mathbb{R}^m} ds \\
&\leq \int_0^{T^+} L(R)\|u(s) - v(s)\|_{\mathbb{R}^m} ds \\
&\leq \|u - v\| \int_0^{T+} L(R)\, ds \\
&= T^+ L(R)\|u - v\| \\
&\leq \frac{1}{2}\|u - v\|.
\end{aligned}$$

□

Remark 18.13. *Having constructed a solution to* (IE)*, it is possible to restart the argument on a second time interval, leading to a sequence of integral*

equations each with an existence and uniqueness property on a specific interval:

$$\begin{aligned}
u(t) &= u_0 + \int_0^t f(u(s))\,ds, && t \in [0, T_1^+] \\
u(t) &= u(T_1^+) + \int_{T_1^+}^t f(u(s))\,ds, && t \in [T_1^+, T_2^+] \\
&\vdots \qquad \vdots \qquad \vdots \\
u(t) &= u(T_j^+) + \int_{T_j^+}^t f(u(s))\,ds, && t \in [T_j^+, T_{j+1}^+].
\end{aligned}$$

Each time interval $|T_{j+1}^+ - T_j^+|$ will depend on $\|u(T_j^+)\|_{\mathbb{R}^m}$. As a consequence, it can happen that $T_j^+ \to T^* < \infty$ as $j \to \infty$; necessarily, then, $u(t) \to \infty$ as $t \to T^*$ from below: if f is locally Lipschitz, a solution to (IE)*, and hence to* (ODE)*, can only cease to exist if the solution becomes unbounded at a finite time.*

18.5 Uniform Contraction and Consequences

Let $(\Theta, \|\cdot\|_\Theta)$ be a Banach space and $T : X \times \Theta \to X$.

Theorem 18.14 (Uniform Contraction Mapping Principle)**.** *Assume that:*

(i) *$T(x, \cdot) : \Theta \to X$ is continuous for each x;*

(ii) *$T(\cdot, \theta) : X \to X$ is a contraction on M for each $\theta \in D$, with D an open, non-empty subset of Θ;*

(iii) *the contraction constant $\lambda(\theta)$ of $T(\cdot, \theta)$ satisfies $\sup_{\theta \in D} \lambda(\theta) \le \alpha < 1$.*

Assume that $T(\cdot, \theta)$ maps M into itself for each $\theta \in D$. Then the equation $T(u, \theta) = u$ has a unique solution $u(\theta) \in M$ for all $\theta \in D$, and $u(\theta)$ is continuous as a function of $\theta \in D$.

Proof Existence and uniqueness follow from the Contraction Mapping Principle. We now fix $\theta \in D$. For any $\varepsilon > 0$, choose $\delta > 0$ such that, for all φ with $\|\theta - \varphi\|_\Theta < \delta$,

$$\|T(u(\theta), \theta) - T(u(\theta), \varphi)\| < (1 - \alpha)\varepsilon.$$

Then

$$\begin{aligned}\|u(\theta) - u(\varphi)\| &= \|T(u(\theta), \theta) - T(u(\varphi), \varphi)\| \\ &\leq \|T(u(\theta), \theta) - T(u(\theta), \varphi)\| + \|T(u(\theta), \varphi) - T(u(\varphi), \varphi)\| \\ &< (1 - \alpha)\varepsilon + \alpha\|u(\theta) - u(\varphi), \|\end{aligned}$$

hence $\|u(\theta) - u(\varphi)\| < \varepsilon$. That is to say, for any $\varepsilon > 0$, there is $\delta > 0$ such that $\|\theta - \varphi\|_\Theta < \delta$ implies $\|u(\theta) - u(\varphi)\| < \varepsilon$, demonstrating that $u : D \subset \Theta \to X$ is continuous. □

Exercises

18.1 Let $(H, \langle\cdot,\cdot\rangle, \|\cdot\|)$ be a real Hilbert space, and let $A \in \mathcal{L}(H, H)$ be such that there exists $\beta > 0$ with $\langle u, Au\rangle \geq \beta\|u\|^2$ for any $u \in H$. Let $\varphi \in H$. For each $\rho > 0$, define the map $T_\rho : H \to H$ by

$$T_\rho u = u - \rho(Au - \varphi).$$

Using the contraction mapping theorem, show that there exists a unique solution to the equation $Au = \varphi$. This provides an alternative proof to part of Theorem 12.12, the Lax–Milgram theorem.

18.2 Let $(X, \|\cdot\|)$ be a Banach space, and let $U \subset X$, $I \subset \mathbb{R}$ be open. Let $f : I \times U \to X$ be such that

- for each $u \in U$, the map $t \mapsto f(t, u)$ is continuous on I;
- for each $t \in I$, the map $u \mapsto f(t, u)$ is Lipschitz continuous on U, with Lipschitz constant independent of t;
- f is bounded on closed bounded subsets of $I \times U$.

For $\varepsilon, r > 0$, define the sets

$$I_\varepsilon = \{t \in I : |t - t_0| \leq \varepsilon\}, \quad U_r = \{u \in U : \|u - u_0\| \leq r\}.$$

Define also the space of continuous functions $u : I_\varepsilon \to U_r$,

$$Z_{\varepsilon,r} = C(I_\varepsilon, U_r),$$

equipped with the norm

$$\|u\|_\varepsilon = \sup_{|t-t_0|\leq\varepsilon} \|u(t)\|.$$

(a) Show that $Z_{\varepsilon,r}$ is complete with respect to $\|\cdot\|_\varepsilon$.

(b) Define the map $P : Z_{\varepsilon,r} \to C(I, U)$ by

$$(Pu)(t) = u_0 + \int_{t_0}^{t} f(s, u(s))\, ds.$$

(i) Show that $P : Z_{\varepsilon,r} \to Z_{\varepsilon,r}$ for appropriate choices of ε and r.

(ii) Show that P is a contraction on $Z_{\varepsilon,r}$ for appropriate choices of ε and r.

(c) Deduce that there exists a unique local solution $u : (t_0 - \varepsilon, t_0 + \varepsilon) \to U$ to the first order ODE

$$\frac{du}{dt}(t) = f(t, u(t)), \quad u(t_0) = u_0.$$

18.3 Prove Lemma 18.3.

19
Implicit Function Theorem

In this lecture we address the following question: if we can solve an equation at one particular value of a parameter which enters the equation, when can we assert that the equation is also solvable for nearby parameter values? This may be addressed in a very clean fashion, with wide-ranging applicability, by a neat application of the contraction mapping principle known as the implicit function theorem.

19.1 Motivation

Consider a parameterized equation of the following form:

$$F(x, \theta) = 0. \tag{19.1}$$

We are interested in finding $x(\theta)$ that solves (19.1) and understanding how the solution varies with θ. Assume that we have (x_0, θ_0) such that

$$F(x_0, \theta_0) = 0.$$

Let $D_x F$ denote the derivative of F with respect to the first variable, and consider the mapping

$$T(x, \theta) = x - D_x F(x_0, \theta_0)^{-1} F(x, \theta). \tag{19.2}$$

Note that any fixed point $x(\theta)$ of T solves $F(x(\theta), \theta) = 0$, provided that $D_x F(x_0, \theta_0)^{-1}$ exists. Moreover,

$$D_x T(x_0, \theta_0) = I - D_x F(x_0, \theta_0)^{-1} D_x F(x_0, \theta_0) = 0.$$

As a consequence, T will be a uniform contraction near to $(x, \theta) = (x_0, \theta_0)$ and Theorem 18.14 can be applied. We will formalize this intuition in section 19.3. But before this, we give a motivating example.

Example 19.1. Consider the following equation for x, parameterized by θ :

$$x^2 - 1 + \theta = 0.$$

The solutions, as a function of θ, are shown in Figure 19.1. The derivative of the left-hand side is $2x$, and when evaluated at either solution $x = \pm\sqrt{1-\theta}$, the derivative is $\pm 2\sqrt{1-\theta}$. For $\theta < 1$ this derivative is non-zero, whilst for $\theta = 1$ the derivative is zero. Note that at any point where $\theta < 1$, so that the derivative is non-zero, either solution may be continued to a neighborhood of any parameter value θ. In contrast, at $\theta = 1$, the solution cannot be continued to a neighborhood of $\theta = 1$: in particular no solution exists for $\theta > 1$.

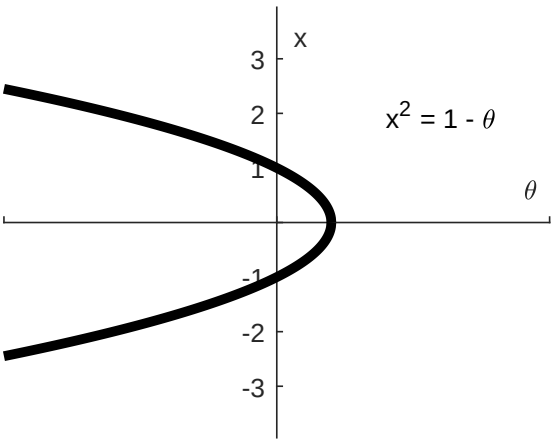


Figure 19.1 An illustration of why the IFT might fail.

Referring back to this example from section 19.3, it will become clear that the implicit function theorem holds everywhere for $\theta < 1$, with the reason for failure at $\theta = 1$ being the zero derivative at that point. ◊

19.2 Mean Value Theorem

Let $(X, \|\cdot\|_X)$, $(Y, \|\cdot\|_Y)$ be Banach spaces, $U \subset X$ bounded and open.

Definition 19.2. *A function $f : U \to Y$ is* Fréchet differentiable *at $x_0 \in U$ if there exists $(Df)(x_0) \in \mathcal{L}(X, Y)$ such that*

$$R(h) := \frac{\|f(x_0 + h) - f(x_0) - (Df)(x_0)h\|_Y}{\|h\|_X}$$

satisfies $R(h) \to 0$ as $h \to 0$ in X. The map $(Df)(x_0)$ is called the Fréchet derivative at the point $x_0 \in U$. We write $f \in C(U, Y)$ if $f : U \to Y$ is Fréchet differentiable at every point in U.

For ease of notation, we will write $(Df)(x_0) = Df(x_0)$.

Remark 19.3. *If $X = \mathbb{R}^n, Y = \mathbb{R}^m$ and f has two derivatives, then*

$$f(x_0 + h) = f(x_0) + Df(x_0)h + O(\|h\|^2_{\mathbb{R}^n}) \qquad \text{(by Taylor theorem)}$$
$$R(h) = O(\|h\|_{\mathbb{R}^n}),$$
$$\underbrace{(Df(x_0))_{ij}}_{\mathbb{R}^{m\times n}} = \frac{\partial f_i}{\partial x_j}(x)\Big|_{x=x_0}$$

Theorem 19.4 (Mean Value Theorem)**.** *If $f \in C(U, Y)$, then*

$$f(x + h) = f(x) + \left(\int_0^1 Df(x + sh)\, ds\right) h.$$

19.3 Implicit Function Theorem

Let X, Y, U be as in the previous section, let $(\Theta, \|\cdot\|_\Theta)$ be a Banach space, and let $D \subset \Theta$ be a bounded, open subset of Θ, and consider equation (19.1). We make the following assumptions concerning F.

Assumptions 19.5.
- $F : U \times D \to X$;
- $F \in C(U \times D, X)$ *with derivative* $\begin{pmatrix} D_x F \\ D_\theta F \end{pmatrix}$;
- $F(x_0, \theta_0) = 0$ *for some* $(x_0, \theta_0) \in U \times D$;
- *there exists* $D_x F(x_0, \theta_0)^{-1} \in \mathcal{L}(X, X)$ *such that*

$$D_x F(x_0, \theta_0)^{-1} D_x F(x_0, \theta_0) = I \in \mathcal{L}(X, X);$$

- $D_x F : U \times D \to \mathcal{L}(X, X)$ *is continuous.*

Example 19.6. If $F : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$ then $D_x F(x_0, \theta_0) \in \mathbb{R}^{n\times n}$. The last two bullet points of Assumption 19.5 will hold if $D_x F(x_0, \theta_0) \in \mathbb{R}^{n\times n}$ is invertible and $D^2 F$ (the second derivative with respect to (x, θ)) exists. ◊

Theorem 19.7 (Implicit Function Theorem)**.** *Under the Assumptions 19.5, there is a neighborhood $U_1 \times D_1$ of (x_0, θ_0), contained in $U \times D \subset X \times \Theta$, and a function $f \in C(D_1, U_1)$, such that*

- $x_0 = f(\theta_0)$;
- $F(x, \theta) = 0$ *for* $(x, \theta) \in U_1 \times D_1$ *if and only if* $x = f(\theta)$.

Proof First, we recall the definition of operator $T(x, \theta)$ in (19.2), and we note that

- $x = T(x, \theta) \quad \Longleftrightarrow \quad F(x, \theta) = 0$;
- $D_x T(x, \theta) = I - D_x F(x_0, \theta_0)^{-1} D_x F(x, \theta)$;
- $D_x T(x_0, \theta_0) = 0$.

By continuity of $D_x F \colon U \times D \to \mathcal{L}(X, X)$, we deduce that there is a neighborhood $U_2 \times D_2 \subseteq U \times D$ of (x_0, θ_0) such that

$$\sup_{(x,\theta)\in U_2\times D_2} \|D_x T(x, \theta)\|_{\mathcal{L}(X,X)} < \frac{1}{2}. \tag{19.3}$$

Choose $U_1 \subseteq U_2$ to be $B(x_0, r)$ (an open ball of radius r centered at x_0) and then choose $D_1 \subseteq D_2$ so that

$$\sup_{\varphi\in D_1} \|T(x_0, \varphi) - x_0\|_X < \frac{r}{2}.$$

Let $M = \overline{B(x_0, r)}$ (closed ball). We prove that $T(\cdot, \theta) \colon M \to M$ and that $T(\cdot, \theta)$ is a uniform contraction on $U_1 \times D_1$. Let $x \in M$ and $\theta \in D_1$. Then

$$\begin{aligned}
\|T(x, \theta) - x_0\|_X &= \|T(x, \theta) - T(x_0, \theta) + T(x_0, \theta) - x_0\|_X \\
&\le \|T(x, \theta) - T(x_0, \theta)\|_X + \|T(x_0, \theta) - x_0\|_X \\
&\le \left\| \left(\int_0^1 D_x T(x + s(x_0 - x), \theta)\, ds \right) (x_0 - x) \right\|_X + \frac{r}{2} \\
&\le \left(\int_0^1 \|D_x T\left((1 - s)x + s x_0, \theta\right)\|_{\mathcal{L}(X,X)}\, ds \right) \|x_0 - x\|_X + \frac{r}{2} \\
&\le \frac{1}{2}\|x_0 - x\|_X + \frac{r}{2} \qquad \text{by (19.3)} \\
&< \frac{r}{2} + \frac{r}{2} = r,
\end{aligned}$$

as required. We now show uniform contraction. Let $x, y \in M$, $\theta \in D_1$.

Then,

$$\begin{aligned}\|T(x,\theta)-T(y,\theta)\|_X &\le \left(\int_0^1 \left\|D_x T\left((1-s)x+sy,\theta\right)\right\|_{\mathcal{L}(X,X)}\,ds\right)\|y-x\|_X\\ &< \frac{1}{2}\|y-x\|_X \qquad \text{by (19.3).}\end{aligned}$$

The uniform contraction mapping theorem gives existence and uniqueness of the solution, concluding the proof. □

Remark 19.8. *Since the neighborhood $U_1 \times D_1$ mentioned in the statement of the theorem is not introduced constructively, one might wonder why Theorem 19.7 would ever fail and, in particular, why it would not be possible to apply the theorem iteratively and cover an arbitrary region in this fashion. Revisiting Example 19.1, it can be argued that the neighborhoods (arising from applying the theorem recursively) can get smaller and smaller, and accumulate on a point, as the parameter approaches a place where the derivative $D_x F$ is not invertible; this is exactly what happens in Example 19.1 at the point where $x = 0, \theta = 1$.*

Example 19.9. Let $A(\theta) \in \mathcal{L}(X,X)$ for each $\theta \in D \subset \mathbb{R}^n$, where D, as above, is bounded and open. Consider the eigenvalue problem of finding $(x,\lambda) \in X \times \mathbb{R}$ that solves

$$\begin{aligned} A(\theta)x &= \lambda x,\\ \|x\|^2 &= 1.\end{aligned} \tag{EVP}$$

Assume $\left(X, \|\cdot\|, \langle\cdot,\cdot\rangle\right)$ is a Hilbert space and $A(\theta)$ is symmetric for every $\theta \in D$: $\langle Au, v\rangle = \langle u, Av\rangle$ for any $u, v \in X$. Assume that for $\theta = 0$, (EVP) has a solution (x_0, λ_0) such that

$$\operatorname{Ker}(A(0) - \lambda_0 I) = \operatorname{span}\{x_0\}.$$

Assume also that $A \in C\left(D, \mathcal{L}(X,X)\right)$. Then, there exists a neighborhood $U_1 \times D_1 \subseteq \left(X \times \mathbb{R}\right) \times \mathbb{R}^n$ of $\left((x_0,\lambda_0),0\right)$ such that (EVP) has a solution, and in particular:

- $\left(x(\theta),\lambda(\theta)\right) \in U_1$ for $\theta \in D_1$;
- $\left(x(0),\lambda(0)\right) = \left(x_0,\lambda_0\right)$;
- the mapping $\theta \mapsto \left(x(\theta),\lambda(\theta)\right)$ is continuous when viewed as a mapping from D_1 into $X \times \mathbb{R}$.

In order to prove this, we first need to reformulate (EVP). To this end, define

$$F(x,\lambda,\theta) \coloneqq \begin{pmatrix} A(\theta)x - \lambda x \\ \frac{1}{2}\|x\|^2 - \frac{1}{2}\end{pmatrix}$$

Solving $F(x, \lambda, \theta) = 0$ for (x, λ) is equivalent to (EVP). We are also given $F(x_0, \lambda_0, 0) = 0$. The derivative with respect to θ, $D_\theta F(x, \lambda, \theta) = D_\theta A(\theta)x$, exists by assumption, and

$$D_{(x,\lambda)}F(x, \lambda, \theta) = \begin{pmatrix} A(\theta) - \lambda I & -x \\ x^* & 0 \end{pmatrix} =: B(x, \lambda, \theta).$$

For $A_0 = A(0)$, $B(x_0, \lambda_0, 0)$ is

$$B(x_0, \lambda_0, 0) = \begin{pmatrix} A_0 - \lambda_0 I & -x_0 \\ x_0^* & 0 \end{pmatrix} \in \mathcal{L}(X \times \mathbb{R}, X \times \mathbb{R}).$$

The result is proved if we show that $B(x_0, \lambda_0, 0)$ is invertible (by applying Theorem 19.7). Thus, it suffices to show that, for any $z \in X, \gamma \in \mathbb{R}$,

$$B(x_0, \lambda_0, 0)\begin{pmatrix} z \\ \gamma \end{pmatrix} = 0 \quad \Rightarrow \quad \begin{pmatrix} z \\ \gamma \end{pmatrix} = 0.$$

Therefore, we study

$$\begin{aligned} (A_0 - \lambda_0 I)z - \gamma x_0 &= 0 \\ \langle x_0, z \rangle &= 0. \end{aligned}$$

Taking the inner product with x_0 in the first equation, we obtain:

$$\begin{aligned} & \left\langle x_0, (A_0 - \lambda I)z \right\rangle - \gamma \|x_0\|^2 = 0 \\ \Rightarrow \quad & \left\langle (A_0 - \lambda I)x_0, z \right\rangle - \gamma \|x_0\|^2 = 0 \\ \Rightarrow \quad & 0 - \gamma \|x_0\|^2 = 0 \\ \Rightarrow \quad & \gamma = 0. \end{aligned}$$

Here we have used that x_0 is non-zero and in the nullspace of $A_0 - \lambda I$. Since $\gamma = 0$, it follows that $(A_0 - \lambda_0 I)z = 0$, and since z is orthogonal to x_0 and the nullspace is only $\operatorname{span}\{x_0\}$, this gives the desired result: $z = 0$. ◊

20
Gradient Descent

Gradient descent is a basic building block for many algorithms in optimization. Its roots are in the continuous time dynamics of a particular class of differential equations, which we present in a general setting. But it is typically implemented through discretization of these equations using a time-stepping method. In this lecture we study three algorithmic choices which arise from this general setting and give supporting theory to inform those choices. Throughout this lecture we will use the following notation for the Euclidean norm and inner product: $\left(\mathbb{R}^n, \langle \cdot, \cdot \rangle, \|\cdot\|\right)$.

20.1 Basic Idea

We focus on minimization of positive function $\Phi : \mathbb{R}^n \to \mathbb{R}^+$. We study algorithms for this problem in general, and in the context of solving linear equations and hence minimization of a quadratic objective function.

Example 20.1. Let $A \in \mathbb{R}^{n\times n}$, $b \in \mathbb{R}^n$ and assume that A is symmetric and strictly positive-definite. Solving the equation

$$Au^* = b$$

to determine the unique solution u^* is then equivalent to minimizing any squared-norm of the vector $Au - b$ over choice of u. In particular, we may introduce

$$\Phi(u) := \frac{1}{2}\|A^{-\frac{1}{2}}(Au - b)\|^2 \tag{20.1}$$

and seek to minimize $\Phi(\cdot)$ in order to find u^*. We note that

$$\Phi(u) = \frac{1}{2}\langle u, Au\rangle - \langle b, u\rangle + \frac{1}{2}\langle b, A^{-1}b\rangle$$

and that

$$\nabla\Phi(u) = Au - b.$$

The Hessian of Φ is

$$D^2\Phi(u) = A.$$

Furthermore note that, since $b = Au^*$,

$$\Phi(u) = \frac{1}{2}\|A^{\frac{1}{2}}(u - u^*)\|^2.$$

◊

For the general setting, let $K \in \mathbb{R}^{n\times n}$ be symmetric and strictly positive-definite and consider the differential equation

$$\frac{du}{dt} = -K\nabla\Phi(u). \tag{20.2}$$

This (in general) nonlinear differential equation has the important property that $\Phi\big(u(t)\big)$ is strictly decreasing unless $u(t)$ is contained in the set of critical points of $\nabla\Phi$. To see this, note that

$$\begin{aligned}\frac{d}{dt}\Big\{\Phi\big(u(t)\big)\Big\} &= \Big\langle \nabla\Phi\big(u(t)\big), \frac{du}{dt}(t)\Big\rangle \\ &= -\Big\langle \nabla\Phi\big(u(t)\big), K\nabla\Phi\big(u(t)\big)\Big\rangle \\ &= -\|K^{\frac{1}{2}}\nabla\Phi\big(u(t)\big)\|^2.\end{aligned}$$

This establishes the desired property. See Section 20.5 for further details.

Making an algorithm out of this property of the differential equation requires discretizing (20.2). To be concrete, we study the family of algorithms (parameterized by θ) given by

$$\frac{u_{n+1} - u_n}{\Delta t_n} = -\theta K\nabla\Phi(u_{n+1}) - (1-\theta)K\nabla\Phi(u_n). \tag{20.3}$$

If we introduce the times $\{t_n\}_{n\in\mathbb{Z}^+}$ given by

$$t_n = t_{n-1} + \Delta t_n, \quad t_0 = 0, \tag{20.4}$$

then we see that $u_n \approx u(t_n)$.

Remark 20.2. *Choice of θ, $\{\Delta t_n\}_{n\in\mathbb{Z}^+}$ and K are design choices which lead to different algorithms; we now discuss these choices in turn.*

20.2 Choice of Discretization

We will concentrate on the choices $\theta \in \{0, 1\}$; note that when $\theta = 0$ the method is *explicit* and does not require the solution of an (in general nonlinear) equation at each step; in contrast, with the choice $\theta = 1$ the method is *implicit* and requires solution of an (in general nonlinear) equation at each step. We will show that the advantage of explicit methods over implicit methods is counter-balanced by desirable properties of implicit methods not shared by explicit methods.

Assumption 20.3. *There exists $\gamma \geq 0$ such that, for all $(u, \xi) \in \mathbb{R}^n \times \mathbb{R}^n$,*

$$\langle \xi, D^2\Phi(u)\xi \rangle \geq -\gamma \|\xi\|^2.$$

This assumption is weaker than convexity of Φ but essentially means that Φ may be rendered convex by addition of the positive quadratic form $\frac{1}{2}\gamma\|u\|^2$.

Example 20.4. Let $\Phi : \mathbb{R} \to \mathbb{R}$ be given by

$$\Phi(u) = \frac{1}{4}(1 - u^2)^2.$$

Then $\Phi'(u) = u^3 - u$ and $\Phi''(u) = 3u^2 - 1$ so that Assumption 20.3 holds with $\gamma = 1$. Note that Φ has three critical points and is not convex; however

$$\Phi(u) + \frac{1}{2}u^2 = \frac{1}{4}(1 + u^4)$$

has one critical point and is convex. ◊

Lemma 20.5. *Under Assumption 20.3 it follows that, for all $(a, b) \in \mathbb{R}^n \times \mathbb{R}^n$,*

$$\Phi(a) - \Phi(b) \leq \langle \nabla\Phi(a), a - b \rangle + \frac{\gamma}{2}\|a - b\|^2.$$

Proof By Taylor series expansion,

$$\begin{aligned}\Phi(b) &= \Phi(a) + \langle \nabla\Phi(a), b - a \rangle + \frac{1}{2}\int_0^1 \langle b - a, D^2\Phi\big(sa + (1 - s)b\big)(b - a)\rangle ds, \\ &\geq \Phi(a) + \langle \nabla\Phi(a), b - a \rangle - \frac{1}{2}\gamma\|b - a\|^2.\end{aligned}$$

Rearranging gives the desired result. □

Proposition 20.6. *Let $\theta = 1$ and set $K = I$. Then provided $\Delta t_n \leq 2/\gamma$, the algorithm* (20.3) *generates a solution sequence satisfying*

$$\Phi(u_{n+1}) \leq \Phi(u_n) \quad \forall n \in \mathbb{Z}^+.$$

Proof Use of the preceding lemma with $a = u_{n+1}$ and $b = u_n$, noting that

$$u_{n+1} - u_n = -\Delta t \nabla \Phi(u_{n+1}),$$

gives

$$\begin{aligned}\Phi(u_{n+1}) - \Phi(u_n) &\le \langle \nabla \Phi(u_{n+1}), u_{n+1} - u_n \rangle + \frac{1}{2}\gamma \|u_{n+1} - u_n\|^2 \\ &\le -\Delta t_n \Big(1 - \frac{1}{2}\gamma \Delta t_n\Big) \|\nabla \Phi(u_{n+1})\|^2.\end{aligned}$$

□

The following proposition shows that, even when $\gamma = 0$ so that the implicit method with $\theta = 1$ suffers no restriction on Δt_n to ensure that $\Phi(u_n)$ is decreasing, the explicit method with $\theta = 0$ suffers a potentially severe restriction to ensure the same.

Proposition 20.7. *Let $\theta = 0$ and set $K = I$ and assume that Φ is given as in Example 20.1 with $b = u^* = 0$:*

$$\Phi(u) = \frac{1}{2}\langle u, Au \rangle. \tag{20.5}$$

If $\inf_{n \in \mathbb{Z}^+} \Delta t_n \ge \Delta t_- > 2/\rho(A)$ for some $\Delta t_- > 0$, then there exists u_0 with the property that the algorithm (20.3) *generates a solution sequence satisfying $\|u_n\| \to \infty$ and $\Phi(u_n) \to \infty$ as $n \to \infty$.*

Proof Let φ be an eigenvector of norm one satisfying $A\varphi = \rho(A)\varphi$. If $u_0 = \varphi$, then a simple induction argument shows that $u_n = \alpha_n \varphi$, where

$$\alpha_{n+1} = \big(1 - \Delta t_n \rho(A)\big)\alpha_n, \quad \alpha_0 = 1.$$

Hence

$$|\alpha_{n+1}| \ge |\Delta t_- \rho(A) - 1||\alpha_n|, \quad |\alpha_0| = 1.$$

By assumption there is $r > 1$ such that

$$|\alpha_{n+1}| \ge r|\alpha_n|, \quad |\alpha_0| = 1.$$

It follows that $|\alpha_n| \to \infty$. Thus

$$\|u_n\| = |\alpha_n| \to \infty, \quad \Phi(u_n) = \frac{1}{2}\alpha_n^2 \rho(A) \to \infty$$

and the desired result follows. □

20.3 Choice of Time-Step

The previous section related to the choice of discretization method through parameter θ, but also demonstrated that restrictions on time-step Δt_n arose naturally from our considerations. We now focus on the explicit method, (20.3) with $\theta = 0$, when applied to quadratic objective function (20.5). We ask if there is an optimal choice of time-step Δt_n.

In the stated setting we have

$$\begin{aligned}\Phi(u_{n+1}) &= \frac{1}{2}\langle u_n - \Delta t_n A u_n, A(u_n - \Delta t_n A u_n)\rangle \\ &= \Phi(u_n) - \Delta t_n \|A u_n\|^2 + \frac{1}{2}\Delta t_n^2 \|A^{\frac{3}{2}} u_n\|^2.\end{aligned}$$

The right-hand side is a concave quadratic in Δt_n and is minimized by the choice of Δt_n which makes the right-hand side stationary:

$$\Delta t_n = \|A u_n\|^2 / \|A^{\frac{3}{2}} u_n\|^2.$$

We thus obtain the optimal gradient descent algorithm, in the case $\theta = 0$ and applied to the quadratic objective function (20.5):

$$u_{n+1} = u_n - \frac{\|A u_n\|^2}{\|A^{\frac{3}{2}} u_n\|^2} A u_n.$$

By writing $u_n \mapsto u_n - A^{-1} b$, we may recover the optimal gradient descent algorithm, in the case $\theta = 0$, when applied to the more general quadratic objective function (20.1):

$$u_{n+1} = u_n - \frac{\|A u_n - b\|^2}{\|A^{\frac{1}{2}}(A u_n - b)\|^2}(A u_n - b).$$

We will derive and study this algorithm in detail in Section 20.6. In particular, we will prove:

Theorem 20.8. *Let u^* solve $A u^* = b$. Then*

$$\|u_n - u^*\|^2 \le \kappa(A) \left(1 - \frac{1}{\kappa(A)}\right)^k \|u_0 - u^*\|^2.$$

In the preceding we have used:

Definition 20.9. *The* condition number *(in the Euclidean norm) is defined as follows:*

$$\kappa(A) = \|A\| \|A^{-1}\| = \frac{\lambda_{\max}}{\lambda_{\min}},$$

where $\lambda_{\max}$ and $\lambda_{\min}$ denote the largest and smallest eigenvalue of matrix A, respectively.

20.4 Choice of Preconditioner

The matrix K is often known as a preconditioner. Replacing A by KA in the previous section, and assuming that A and K commute so that KA is symmetric, we deduce that the optimal choice of K is A^{-1}: in Theorem 20.8 this gives convergence in one step. This, of course, is not surprising and somewhat unrealistic: if A^{-1} is known, then of course $Au^* = b$ is easily solved in one step. However, the discussion points to interesting practical preconditioners which, for non-quadratic objective functions, seek to find $K \approx D^2\Phi(\cdot)^{-1}$, and for interacting particle optimization such K can be estimated from the interacting agents.

20.5 Continuous Time

Consider the problem of minimizing $\Phi\colon \mathbb{R}^n \to \mathbb{R}$. We will assume it satisfies the following:

Assumptions 20.10.
- $\nabla\Phi$ *is locally Lipschitz;*
- $\Phi(u) \geq 0 \quad \forall u \in \mathbb{R}^n$*;*
- $\forall \varphi \geq 0$*, there exists a function* $R = R(\varphi)$ *such that* $\Phi(u) \leq \varphi \implies u \in B(0, R)$.

Consider the ODE

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\nabla\Phi(u), \quad u(0) = u_0. \tag{GD}$$

Theorem 20.11. *Under Assumptions 20.10, the initial-value problem* (GD) *has a unique solution* $u \in C\big([0,\infty), \mathbb{R}^n\big)$ *satisfying*

- $\Phi(u(t))$ *is non-increasing in* t*;*
- $\Phi(u(t)) \to \Phi^*$ *as* $t \to \infty$*;*
- *if* $u(t_j) \to u^*$ *as* $t_j \to \infty$*, then* $\Phi(u^*) = \Phi^*$.

Proof We first assume a solution exists; then,

$$\frac{\mathrm{d}}{\mathrm{d}t}\big(\Phi(u(t))\big) = \left\langle \nabla\Phi(u(t)), \frac{\mathrm{d}u}{\mathrm{d}t}(t) \right\rangle = -\left\| \frac{\mathrm{d}u}{\mathrm{d}t}(t) \right\|^2 \leq 0. \tag{20.6}$$

Therefore, $\Phi(u(t_2)) \leq \Phi(u(t_1))$ if $t_2 \geq t_1 \geq 0$ (i.e., non-increasing).

(i) Define $R^+ = R(\Phi(u_0))$. By the last bullet in Assumptions 20.10 and using the fact that $\Phi(u(t)) \leq \Phi(u_0)$, we conclude that, whilst a solution exists, $\|u(t)\| \leq R^+$ for all $t \geq 0$.

(ii) By Theorem 18.12, (IE-GD) has a unique solution in a ball around u_0 on time interval $[0, T^+]$:

$$u(t) = u_0 - \int_0^t \nabla\Phi(u(s))\,\mathrm{d}s. \tag{IE-GD}$$

Examination of the contraction mapping argument proving existence of this solution demonstrates that T^+ depends on u_0 only through its norm.

(iii) We can repeat the local existence argument on $[T^+, 2T^+]$, $[2T^+, 3T^+]$, ... because (20.6) implies that $\|u(jT^+)\| \leq R^+$ so that intervals of the same length can be used (in contrast to the general situation outlined in Remark 18.13).

(iv) The solution to (IE-GD) is continuously differentiable and is hence also a solution to (GD). Finally, the solution is unique not just in a ball around u_0, but globally in $C\big([0,\infty);\mathbb{R}^n\big)$. This may be proved by changing Φ outside $B(0, 2R^+)$ to make $\nabla\Phi$ globally Lipschitz and then proving uniqueness for this problem. Since the solution, if it exists, remains in $\overline{B(0, R^+)}$, the modification outside $B(0, 2R^+)$ has no bearing on the uniqueness conclusion.

Since $\Phi(u(t))$ is non-increasing and bounded from below, by the monotone convergence theorem we have that $\Phi(u(t)) \to \Phi^* \geq 0$ as $t \to \infty$. Let $\{t_j\}_{j\in\mathbb{N}}$ be such that $u(t_j) \to u^*$. Then

$$\begin{aligned}
&|\Phi(u^*) - \Phi^*| \leq |\Phi(u^*) - \Phi(u(t_j))| + |\Phi(u(t_j)) - \Phi^*| \\
&\leq \left(\int_0^1 \|\nabla\Phi(u^* + s(u(t_j) - u^*))\|\,\mathrm{d}s\right) \|u(t_j) - u^*\| + |\Phi(u(t_j)) - \Phi^*| \\
&\leq C\|u(t_j) - u^*\| + |\Phi(u(t_j)) - \Phi(u^*)|,
\end{aligned}$$

and both terms tend to 0 as $j \to \infty$, thus, $\Phi(u^*) = \Phi^*$.

□

Remark 20.12. *The preceding result shows that* Φ *decreases along trajectories of the gradient flow* (GD) *and is hence suggestive that it is useful for purposes of optimization. In fact, if the critical points (maxima, saddles, minima) of* Φ *are isolated, then* u^* *must be a critical point (i.e.,* $\nabla\Phi(u^*) = 0$*). Furthermore, if*

the initial condition is chosen at random with respect to a probability measure which has density with respect to Lebesgue measure, then it is a straightforward consequence that, with probability one, the solution will converge to a (possibly local) minimum; however, ensuring that the global minimum is reached requires new ideas, such as the addition of noise, the use of multiple starting points u_0 and mini-batching of Φ.

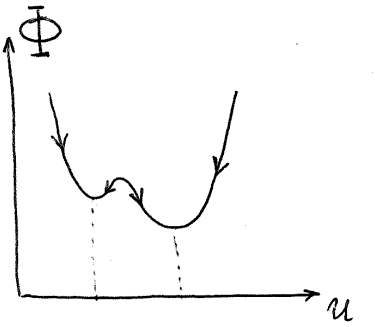


Figure 20.1 An illustration of the GD algorithm for a function with two local minima.

20.6 Discrete Time

Here we consider the gradient flow (GD) discretized as in (20.3) with $\theta = 0$, the forward Euler discretization:

$$u_{k+1} = u_k - \alpha_k \nabla \Phi(u_k). \tag{20.7}$$

The choice of time-step α_k will then affect the performance of the algorithm. In order to get insight into this choice, we study it in this section for the specific case of a quadratic function Φ.

Let $A \in \mathbb{R}^{n\times n}$ be (strictly) positive-definite and symmetric. Then solving $Au = b$ is equivalent to minimizing $\|Au - b\|^2_{\mathbb{R}^n}$ in any norm on $\mathbb{R}^n$. We use this connection, for a specific choice of norm, in what follows. To this end the following definition is useful:

Definition 20.13. *For A satisfying assumptions above, define the weighted inner product*

$$\langle u, v\rangle_A := \langle u, Av\rangle = \langle A^{1/2}u, A^{1/2}v\rangle$$

and the weighted norm

$$\|u\|_A := \left(\|A^{1/2}u\|^2\right)^{1/2}.$$

Remark 20.14. *$A^{1/2}$ is well-defined because A is symmetric positive-definite, and hence it is diagonalizable by a matrix of eigenvectors. Indeed if $A = P\Lambda P^\top$, where Λ is diagonal with positive diagonal entries comprising eigenvalues, then $A^{1/2} := P\Lambda^{1/2}P^\top$.*

For $A \in \mathbb{R}^{n\times n}$ symmetric positive-definite, define

$$\Phi(u) = \frac{1}{2}\|Au - b\|^2_{A^{-1}}, \tag{20.8}$$

noting that this coincides with Φ as defined in (20.1).

Lemma 20.15. *The function $\Phi(u)$ defined above can be re-expressed in any of the following ways:*

$$\begin{aligned}\Phi(u) &= \frac{1}{2}\|u - A^{-1}b\|^2_A \\ &= \frac{1}{2}\|A^{-1/2}(Au - b)\|^2 \\ &= \frac{1}{2}\langle u, Au\rangle - \langle b, u\rangle + \frac{1}{2}\|b\|^2_{A^{-1}}.\end{aligned}$$

Therefore, $\nabla\Phi(u) = Au - b$. Hence the unique equilibrium point of (GD) *with this Φ is the unique solution u^* of the linear system $Au^* = b$.*

Proof For the first part, write by definition of $\Phi(u)$

$$\begin{aligned}\Phi(u) = \frac{1}{2}\left\langle Au - b, A^{-1}(Au - b)\right\rangle &= \frac{1}{2}\left\langle A(u - A^{-1}b), u - A^{-1}b\right\rangle \\ &= \frac{1}{2}\left\|u - A^{-1}b\right\|^2_A.\end{aligned}$$

Similarly for the second part,

$$\begin{aligned}\Phi(u) &= \frac{1}{2}\left\langle Au - b, A^{-1/2}A^{-1/2}(Au - b)\right\rangle \\ &= \frac{1}{2}\left\langle A^{-1/2}(Au - b), A^{-1/2}(Au - b)\right\rangle \\ &= \frac{1}{2}\left\|A^{-1/2}(Au - b)\right\|^2.\end{aligned}$$

And also,

$$\begin{aligned}\Phi(u) &= \frac{1}{2}\left\langle (Au-b), u - A^{-1}b)\right\rangle \\ &= \frac{1}{2}\langle Au, u\rangle - \frac{1}{2}\langle b, u\rangle - \frac{1}{2}\left\langle Au, A^{-1}b\right\rangle + \frac{1}{2}\|b\|^2_{A^{-1}} \\ &= \frac{1}{2}\langle u, Au\rangle - \frac{1}{2}\langle b, u\rangle - \frac{1}{2}\left\langle u, AA^{-1}b\right\rangle + \frac{1}{2}\|b\|^2_{A^{-1}} \\ &= \frac{1}{2}\langle u, Au\rangle - \langle b, u\rangle + \frac{1}{2}\|b\|^2_{A^{-1}}.\end{aligned}$$

So, finally, finding a critical point u^* of Φ amounts to solving

$$\nabla\Phi(u^*) = Au^* - b = 0 \quad \Longleftrightarrow \quad u = A^{-1}b.$$

□

We now study (20.7) with $\Phi(u)$ given by equation (20.8):

$$\begin{aligned}u_{k+1} &= u_k - \alpha_k(Au_k - b) \\ &= u_k + \alpha_k r_k,\end{aligned}$$

where we have introduced the residual $r(u) = b - Au$ and its discretized version $r_k = b - Au_k = r(u_k)$. We address the question of how to choose α_k in order to maximize the decrease in $\Phi(u_j)$ as index j is updated from k to $k+1$.

To this end, we observe that

$$\begin{aligned}\Phi(u_{k+1}) &= \frac{1}{2}\|Au_k - b + \alpha_k Ar_k\|^2_{A^{-1}} = \frac{1}{2}\|\alpha_k Ar_k - r_k\|^2_{A^{-1}} \\ &= \frac{1}{2}\alpha_k^2\|Ar_k\|^2_{A^{-1}} - \alpha_k\langle Ar_k, r_k\rangle_{A^{-1}} + \frac{1}{2}\|r_k\|^2_{A^{-1}}\end{aligned}$$

and, simply from the definition of r_k, $\Phi(u_k) = \frac{1}{2}\|r_k\|^2_{A^{-1}}$. Hence,

$$\Phi(u_{k+1}) - \Phi(u_k) = \frac{1}{2}\alpha_k^2\|r_k\|^2_A - \alpha_k\|r_k\|^2.$$

The α_k which minimizes the right-hand side (maximizes the decrease $\Phi(u_k) \mapsto \Phi(u_{k+1})$) solves

$$0 = \alpha_k\|r_k\|^2_A - \|r_k\|^2 \quad \Longrightarrow \quad \alpha_k = \frac{\|r_k\|^2}{\|r_k\|^2_A}.$$

So, the algorithm is

$$u_{k+1} = u_k + \frac{\|r_k\|^2}{\|r_k\|^2_A} r_k, \tag{20.9}$$

and hence, the following property holds:

$$\Phi(u_{k+1}) = \Phi(u_k) - \frac{1}{2}\frac{\|r_k\|^4}{\|r_k\|_A^2}. \tag{20.10}$$

Since $A \in \mathbb{R}^{n\times n}$ is symmetric positive-definite, we have the eigenvalue problem

$$A\varphi^{(j)} = \lambda_j \varphi^{(j)}, \quad \langle \varphi^{(i)}, \varphi^{(j)} \rangle = \delta_{ij}$$

and

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n = \lambda_{\min} > 0.$$

We are now in a position to prove Theorem 20.8:

Proof of Theorem 20.8 Expand r_k in an orthonormal basis:

$$r_k = \sum_{j=1}^{n} a_j \varphi^{(j)}, \quad a_j = \langle \varphi^{(j)}, r_k \rangle.$$

Then,

$$\|r_k\|^2 = \sum_{j=1}^{n} a_j^2 \geq \lambda_{\min} \sum_{j=1}^{n} \frac{a_j^2}{\lambda_j} = \lambda_{\min} \|r_k\|_{A^{-1}}^2,$$

$$\|r_k\|^2 = \sum_{j=1}^{n} a_j^2 \geq \frac{1}{\lambda_{\max}} \sum_{j=1}^{n} \lambda_j a_j^2 = \frac{1}{\lambda_{\max}} \|r_k\|_A^2.$$

Hence

$$\frac{1}{2}\frac{\|r_k\|^4}{\|r_k\|_A^2} = \frac{1}{2}\frac{\|r_k\|^2}{\|r_k\|_A^2}\|r_k\|^2 \geq \frac{1}{2}\frac{1}{\lambda_{\max}}\lambda_{\min}\|r_k\|_{A^{-1}}^2.$$

Recalling that $\Phi(u_k) = \frac{1}{2}\|r_k\|_{A^{-1}}^2$, we rewrite the previous inequality:

$$\frac{1}{2}\frac{\|r_k\|^4}{\|r_k\|_A^2} \geq \frac{1}{2}\frac{1}{\kappa(A)}\|r_k\|_{A^{-1}}^2 = \frac{1}{\kappa(A)}\Phi(u_k)$$

By (20.10) we therefore have

$$\Phi(u_{k+1}) \leq \left(1 - \frac{1}{\kappa(A)}\right)\Phi(u_k).$$

By induction it then follows that

$$\Phi(u_k) \leq \left(1 - \frac{1}{\kappa(A)}\right)^k \Phi(u_0). \tag{20.11}$$

By means of orthogonal expansions, such as those above, we have

$$\lambda_{\min}\|v\|^2 \leq \|v\|_A^2 \leq \lambda_{\max}\|v\|^2$$

for any $v \in \mathbb{R}^n$. We can then write

$$\Phi(u) = \frac{1}{2}\left\|u - A^{-1}b\right\|_A^2 \leq \frac{\lambda_{\max}}{2}\left\|u - A^{-1}b\right\|^2, \tag{20.12}$$

$$\Phi(u) = \frac{1}{2}\left\|u - A^{-1}b\right\|_A^2 \geq \frac{\lambda_{\min}}{2}\left\|u - A^{-1}b\right\|^2, \tag{20.13}$$

and noting that $u = A^{-1}b$ and combining (20.11), (20.12), (20.13), we obtain

$$\frac{\lambda_{\min}}{2}\|u_k - u\|^2 \leq \left(1 - \frac{1}{\kappa(A)}\right)^k \frac{\lambda_{\max}}{2}\|u_0 - u\|^2$$

and division by $\lambda_{\min}/2$ gives the desired result.

□

Exercises

20.1 Let $A \in \mathbb{C}^{m\times n}$ and $b \in \mathbb{C}^m$. Define the least squares functional $\Phi : \mathbb{C}^n \to \mathbb{R}$ by

$$\Phi(x) = \frac{1}{2}\|Ax - b\|_2^2.$$

We say that $x \in \mathbb{C}^n$ is a solution to the least squares problem for Φ if

$$\Phi(x) \leq \Phi(y) \quad \text{for all } y \in \mathbb{C}^n. \tag{LSQ}$$

(a) Show that $x \in \mathbb{C}^n$ is a solution to (LSQ) if and only if it satisfies the normal equations

$$A^*Ax = A^*b.$$

(b) When is there a unique solution to (LSQ)?

(c) Show that the vector $A^\dagger b$ solves (LSQ), where $A^\dagger$ denotes the Moore-Penrose pseudoinverse to A. Show also that amongst all solutions to (LSQ), it has minimal 2-norm.

Define now the *Tikhonov regularized* least squares functional $\Phi_\lambda : \mathbb{C}^n \to \mathbb{R}$,

$$\Phi_\lambda(x) = \Phi(x) + \frac{\lambda}{2}\|x\|_2^2,$$

where $\lambda > 0$ is a scalar penalty parameter.

(d) Derive the normal equations that are equivalent to solving (LSQ) for Φ_λ.

(e) Show that for each $\lambda > 0$, there exists a unique $x_\lambda \in \mathbb{C}^n$ solving (LSQ) for Φ_λ.

(f) By considering the SVD of A or otherwise, show that

$$A^\dagger b = \lim_{\lambda \downarrow 0} x_\lambda.$$

20.2 Given a matrix $A \in \mathbb{C}^{m\times n}$, define its condition number by

$$\kappa(A) = \|A\|_2 \|A^\dagger\|_2.$$

Let $b \in \mathbb{C}^m$, and denote by Φ the (non-regularized) least squares functional from Exercise 20.1. Assume that we have an approximation b_ε of b leading to an approximate least squares functional Φ_ε,

$$\Phi_\varepsilon(x) = \frac{1}{2}\|Ax - b_\varepsilon\|_2^2.$$

Let x, x_ε denote the solution to (LSQ) using Φ, Φ_ε, respectively. Define $\theta \in [0, \pi/2]$ via $\cos(\theta) = \|Ax\|_2/\|b\|_2$. Show that the relative error in x_ε satisfies

$$\frac{\|x - x_\varepsilon\|_2}{\|x\|_2} \leq \frac{\kappa(A)}{\eta \cos(\theta)} \frac{\|b - b_\varepsilon\|}{\|b\|},$$

where $\eta = \|A\|_2\|x\|_2/\|Ax\|_2$. Which properties of A, b would lead to high sensitivity of the relative error in x_ε with respect to the relative error in b_ε? What is an explanation for this?

References

Adams, Robert A, and Fournier, John JF. 2003. *Sobolev spaces*. Vol. 140. Academic press.

Evans, Lawrence C. 2002. *Partial differential equations*, Graduate Studies in Mathematics. Vol. 19. American Mathematical Society.

Evans, L.C. 2005. *An Introduction to Mathematical Optimal Control Theory*. UC Berkeley. https://math.berkeley.edu/~evans/control.course.pdf.

Griffel, David H. 2002. *Applied functional analysis*. Courier Corporation.

Hanke, Martin. 2017. *A Taste of Inverse Problems: Basic Theory and Examples*. SIAM.

Hutson, Vivian, Pym, J, and Cloud, M. 2005. *Applications of functional analysis and operator theory*. Vol. 200. Elsevier.

Luenberger, David G. 1969. *Optimization by vector space methods*. John Wiley & Sons.

Robinson, James C. 2020. *An Introduction to Functional Analysis*. Cambridge University Press.

Zeidler, Eberhard. 1995. *Applied functional analysis, Main principles and their application*. Contents of AMS.

Zelnik-Manor, Lihi, and Perona, Pietro. 2004 (01). Self-tuning spectral clustering. Pages 1601–1608 of: *Proceedings of the 17th International Conference on Neural Information Processing Systems*. NIPS'04.