LINEAR VECTOR SPACES
&
APPLICATIONS

by Prof. Stephen Pankavich

$\mathcal{SP}$

Department of

Applied Mathematics and Statistics

Colorado School of Mines

2020

This work was supported by the Colorado Department of Higher Education's Open Educational Resources Project.

http://masterplan.highered.colorado.gov/oer-in-colorado/

Stephen Pankavich
Department of Applied Mathematics and Statistics
Colorado School of Mines
Golden, CO 80401, USA
pankavic (at) mines dot edu

Contents

Chapter 1

Preface

These notes serve to form the foundational knowledge for MATH 500: Linear Vector Spaces at the Colorado School of Mines. The information herein has been compiled, thus far, over three different iterations of the class during the Fall 2014, Fall 2017, and Fall 2018 semesters. The main goals of this course are to

- solidify students' prior knowledge base concerning vector spaces, linear transformations, matrix decompositions, and the relationships between them, while unifying their disparate backgrounds in undergraduate mathematics, like Linear Algebra, Proofs, and Analysis

- increase the level of abstraction within previous linear algebra courses so that results concerning finite-dimensional spaces and subspaces can be understood in a generalized context regarding infinite-dimensional spaces and linear operators defined on them, in addition to generalized notions of symmetry, orthogonality, and duality.

- introduce incoming M.S. and Ph.D. students in both the Statistics and Computational and Applied Mathematics graduate programs to rigorous mathematical proof, while remaining accessible to numerous students from other programs at Mines, including Geophysics, Hydrology, Petroleum Engineering, Electrical Engineering, Computer Science, and Physics, among others.

- focus on specific applications of the mathematical material and motivate creative interest in their extension or application to other fields

- prepare students for MATH 550: Computational Linear Algebra, a follow-up course offered in the Spring semester and focused on the development, implementation, and numerical analysis of algorithms used to (i) solve linear systems of algebraic equations via direct and iterative methods, (ii) approximate eigenvalues, and (iii) compute the QR Factorization and Singular Value Decomposition of a matrix, often for reduced rank approximations, data compression, or for solving least squares problems.

The structure of the material is meant to first introduce abstract concepts, then understand the implications of certain definitions and theorems to finite-dimensional vectors space and matrices, and finally, focus on a useful application of these results. For all of these reasons, we have grouped portions of the material into chapters that

separately focus on theoretical (i.e., theorem-proof based) results and applications of the theory.

Of course, no single source of material will likely encapsulate all the knowledge that we have gained to date regarding Linear Algebra, Vector Spaces, or any of their numerous applications. For this reason we include a list of other materials, mostly textbooks, lecture notes, and review problem sets, that may shed further light on the subject. These include some of the references [1, 11, 16, 27, 31] listed in the bibliography. Additionally, we would like to acknowledge the contributions of Jake Chambers and Brett Powers for assisting with the typesetting of a portion of the notes and thank Prof. Rebecca Swanson for helpful comments and feedback.

In the interest of course replication or enhancement, an outline for MATH 500 is presented below.

Course Outline

1. Introduction and Notation
2. Review of Linear Algebra ($\approx$ 1 week)
    - Basic Definitions
    - Solvability & Invertible Matrix Theorem
    - New definitions & terminology - e.g., Hermitian
3. Application: Google's PageRank algorithm ($\approx$ 1 week)
4. Linear Vector Spaces ($\approx$ 5 weeks)
    - Definition & subspaces
    - Span, linear independence, basis, dimension
    - Properties and examples of finite and infinite dimensional vector spaces
    - Analysis on vector spaces - metrics, norms, inner products & completeness
    - Orthogonality, projections, Gram-Schmidt, and QR factorization
    - Properties of Hilbert spaces - projection and decomposition
5. Application: Linear Regression and Ranking Systems ($\approx$ 1 week)
6. Linear Operators ($\approx$ 3 weeks)
    - Kernel and Image of linear operators
    - Rank-Nullity Theorem
    - Adjoint operator
    - Fundamental Theorem of Linear Algebra
    - Duality
    - Operator and matrix norms
7. Decompositions and Factorizations ($\approx$ 4 weeks)
    - Eigenspaces
    - Operator and matrix diagonalization, Jordan form, and Schur form
    - Unitary, Orthogonal, Hermitian, and Normal operators & matrices
    - Spectral Theorem
    - Singular Value Decomposition - definition, construction, properties
8. Application: Principal Components Analysis ($\approx$ 1 week)

Other Possible Application Topics:

- Markov Chains
- Linear System Applications
  (Circuits, heat & stress distribution, chemical equilibrium, coupled oscillators)
- Computer Graphics

Chapter 2

Introduction & Review Exercises

2.1 Notation

Throughout, we will make a few standard assumptions regarding notation:

- We let $p, q, r \in \mathbb{N}$ be given to represent dimensions of vector spaces or matrices when necessary.
- In general, we will not use vector or boldface notation (i.e., $\vec{v}$ or $\mathbf{v}$) to distinguish vectors from scalars or matrices. Instead, the reader will need to infer the type of object being discussed from context or a statement such as $v \in \mathbb{R}^p$.
- Additionally, all vectors $v \in \mathbb{R}^p$ will be treated as column vectors so that v^T is a row vector.
- The notation rref(A) stands for the "Reduced Row Echelon Form of A"

Additionally, we will assume basic knowledge of Linear Algebra (such as the material discussed in MATH 332: Linear Algebra, here at Mines), including fundamental matrix theory and notation, the Invertible Matrix Theorem, a familiarity with rank(A), Col(A), Nul(A), and eigenvalues and eigenvectors (definition, computing them, multiplicities, etc.). That being said, we will generally not rely upon knowledge of the determinant. Though utilizing the determinant to perform routine calculations by hand was useful in a Linear Algebra setting, it is unnecessary to describe the theory of vector spaces or linear transformations, and it is *rarely* used to determine information about a matrix in a computational framework (in particular, algorithms to compute the determinant of a matrix are not especially fast and do not provide more information than other operations).

2.2 Essential Theorems from Linear Algebra

1. **Existence and Uniqueness Theorem**

   A linear system is consistent if and only if the rightmost column of the augmented matrix is not a pivot a column–that is if and only if an echelon form of the augmented matrix has no row of the form $\left[\begin{array}{ccc|c} 0 & \cdots & 0 & b \end{array}\right]$ with $b \neq 0$. If a linear system is consistent, then the solution set contains either (i) a unique

solution, when there are no free variables, or (ii) infinitely many solutions, when there is at least one free variable.

2. **Theorem 4 Chapter 1**[†] (**Row-Pivot Theorem**)

   Let A be a $p \times q$ matrix. Then the following statements are equivalent:

   (a) For each $b \in \mathbb{R}^p$, the equation $Ax = b$ has a solution (i.e., is consistent).

   (b) Each $b \in \mathbb{R}^p$ is a linear combination of the columns of A.

   (c) The columns of A span $\mathbb{R}^p$.

   (d) A has a pivot position in every row.

3. **Column-Pivot Theorem**

   Let A be a $p \times q$ matrix. Then the following statements are equivalent:

   (a) For each $b \in \mathbb{R}^p$, the equation $Ax = b$ has at most one solution.

   (b) The equation $Ax = 0$ has exactly one solution, namely $x = 0$.

   (c) The columns of A are linearly independent.

   (d) A has a pivot position in every column.

4. **Theorem 5 Chapter 2** (Invertible implies unique solution)

   If A is an invertible $p \times p$ matrix, then for each $b \in \mathbb{R}^p$, the equation $Ax = b$ has the unique solution $x = A^{-1}b$.

5. **Theorem 6 Chapter 2** (Basic Inverse Properties)

   (a) If A is invertible, then A^{-1} is invertible and $(A^{-1})^{-1} = A$.

   (b) If A and B are $p \times p$ invertible matrices, then so is AB, and the inverse of AB is the product of the inverses of A and B in reverse order. That is $(AB)^{-1} = B^{-1}A^{-1}$.

   (c) If A is invertible, then so is A^T, and the inverse of A^T is the transpose of A^{-1}. That is $(A^T)^{-1} = (A^{-1})^T$.

6. **Theorem 1 Chapter 4** (Span is a Subspace)

   If $v_1, \ldots, v_p$ are in a vector space V, then $\text{span}(v_1, \ldots, v_p)$ is a subspace of V.

7. **Theorem 5 Chapter 4** (Spanning Set Theorem) Let $S = \{v_1, \ldots, v_p\}$ be a set in V, and let $H = \text{span}(v_1, \ldots, v_p)$.

   (a) If one of the vectors in S, say v_k, is a linear combination of the remaining vectors in S, then the set formed from S by removing v_k still spans H.

   (b) If $H \neq \{0\}$, then some subset of S is a basis for H.

[†]All Theorem and Chapter numberings refer to *Linear Algebra and Its Applications* by Lay, 4th edition [11].

8. **Theorem 12 Chapter 4** (Basis Theorem) Let V be a p-dimensional vector space with $p \geq 1$. Any linearly independent set of exactly p elements in V is automatically a basis. Any set of exactly p elements that spans V is automatically a basis for V. If a vector space V has a basis of q vectors, then every basis of V must contain exactly q vectors.

9. **Theorem 14 Chapter 4** (Rank-Nullity Theorem) The dimensions of the column space and the row space of a $p \times q$ matrix A are equal. The common dimension, the rank of A, also equals the number of pivot positions in A and satisfies the equation

$$\text{rank}(A) + \dim(\text{Nul}(A)) = q.$$

10. **Theorem 10 Chapter 6** (Orthonormal Columns and Projections)

If $\{u_1, \ldots, u_n\}$ is an orthonormal basis for a subspace $W \subseteq \mathbb{R}^p$, then

$$\text{proj}_W y = (y \cdot u_1)u_1 + \cdots + (y \cdot u_n)u_n.$$

If $U = [u_1 \cdots u_n]$, then

$$\text{proj}_W y = UU^T y$$

for all $y \in \mathbb{R}^p$.

11. **Invertible Matrix Theorem (IMT)** Let A be a $p \times p$ matrix. Then the following statements are equivalent:

(a) A is invertible (or nonsingular).

(b) A is row equivalent to $\mathbb{I}_p$.

(c) A has p pivots.

(d) The equation $Ax = 0$ has only the trivial solution.

(e) The columns of A form a linearly independent set.

(f) The linear transformation $x \to Ax$ is one-to-one.

(g) The equation $Ax = b$ has at least one solution for every b in $\mathbb{R}^p$.

(h) The columns of A span $\mathbb{R}^p$.

(i) The linear transformation $x \to Ax$ is onto.

(j) There is an $p \times p$ matrix C such that $CA = \mathbb{I}_p$.

(k) There is an $p \times p$ matrix D such that $AD = \mathbb{I}_p$.

(l) A^T is invertible.

(m) The columns of A form a basis of $\mathbb{R}^p$.

(n) $\text{Col}(A) = \mathbb{R}^p$ or $\dim(\text{Col}(A)) = p$ or $\text{rank}(A) = p$

(o) $\text{Nul}(A) = \{0\}$ or $\dim(\text{Nul}(A)) = 0$

(p) $\lambda = 0$ is not an eigenvalue of A.

(q) The determinant of A is nonzero.

12. **LU Factorization** (Gaussian Elimination)

    For any $p \times p$ matrix A, there is a $p \times p$ permutation matrix P, an upper triangular $p \times p$ matrix U, and a lower triangular $p \times p$ matrix L with $\ell_{kk} = 1$ for every $k = 1, ..., p$ such that

    $$PA = LU.$$

13. **Spectral Theorem for real, symmetric matrices**

    For any real, $p \times p$ matrix A, A is symmetric if and only if A is orthogonally diagonalizable.

Note that result #2 above concerns $A \in \mathbb{R}^{p \times q}$, while the IMT and LU Factorization apply only to square matrices, i.e. $A \in \mathbb{R}^{p \times p}$.

2.3 Review Exercises

Definition 2.1. Given $A \in \mathbb{C}^{p\times q}$, the **Hermitian transpose** of A, written A^H (or $A^\dagger$ in some references) is defined by

$$A^H = \overline{A}^T = \overline{A^T}.$$

Definition 2.2. We say $A \in \mathbb{C}^{p\times p}$ is **Hermitian** if

$$A^H = A.$$

Note that a Hermitian matrix is a generalization of a symmetric matrix. However, this is different from a complex-valued symmetric matrix, as shown in the following example.

Example 1. Let $A_1 = \begin{bmatrix} 1 & 1+i \\ 1+i & 1 \end{bmatrix}$ and $A_2 = \begin{bmatrix} 1 & 1+i \\ 1-i & 1 \end{bmatrix}$.

1. $A_1^H = \begin{bmatrix} 1 & 1-i \\ 1-i & 1 \end{bmatrix} \neq A_1$ but $A_1^T = A_1$ so A_1 is not Hermitian, but is a complex-valued symmetric matrix.
2. $A_2^H = \begin{bmatrix} 1 & 1+i \\ 1-i & 1 \end{bmatrix} = A_2$ but $A_2^T = \begin{bmatrix} 1 & 1-i \\ 1+i & 1 \end{bmatrix} \neq A_2$.
   So A_2 is Hermitian, but not symmetric.

Theorem 2.1. For $A \in \mathbb{C}^{p\times q}$ and $B \in \mathbb{C}^{q\times r}$, we have the identity

$$(AB)^H = B^H A^H.$$

Proof. Recall that for any $C \in \mathbb{C}^{p\times q}$ and $D \in \mathbb{C}^{q\times r}$

$$[CD]_{ik} = \sum_{j=1}^{q} C_{ij} D_{jk}$$

for $i = 1, ..., p$ and $k = 1, ..., r$. Thus, for A and B given in the theorem

$$\left[(AB)^H\right]_{ik} = \left[\overline{AB}\right]_{ki} = \sum_{j=1}^{q} \overline{A}_{kj}\overline{B}_{ji} \tag{2.1}$$

where the first equality is due to the definition of transpose. Furthermore, we have

$$\begin{aligned} \left[B^H A^H\right]_{ik} &= \sum_{j=1}^{q} \left[\overline{B}^T\right]_{ij} \left[\overline{A}^T\right]_{jk} \\ &= \sum_{j=1}^{q} \overline{B}_{ji}\overline{A}_{kj} \\ &= \sum_{j=1}^{q} \overline{A}_{kj}\overline{B}_{ji}. \end{aligned} \tag{2.2}$$

As (2.1) and (2.2) are equal for all $i = 1, ..., p$ and $k = 1, ..., r$, the result follows. □

Corollary 2.1. For $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{q\times r}$, we have the identity

$$(AB)^T = B^T A^T.$$

Theorem 2.2. Let $A \in \mathbb{C}^{p\times p}$ and $B \in \mathbb{C}^{p\times p}$ be Hermitian matrices. Then AB is Hermitian if and only if A and B commute (i.e. $AB = BA$).

Proof. The proof is left as a homework exercise (cf. Problem 2.3) □

Before we state the next result, recall the following definition.

Definition 2.3. A matrix $A \in \mathbb{R}^{p\times p}$ is **nonsingular** (invertible) if there exists $X \in \mathbb{R}^{p\times p}$ such that $AX = XA = \mathbb{I}_p$. Otherwise, we say A is **singular**.

Theorem 2.3. Let $A \in \mathbb{R}^{p\times p}$ be given. Then, $AX = \mathbb{I}_p$ for some $X \in \mathbb{R}^{p\times p}$ implies $XA = \mathbb{I}_p$

Comment. This theorem can be made into an if and only if statement as well.

Proof. We let the matrices $A \in \mathbb{R}^{p\times p}$ and $X \in \mathbb{R}^{p\times p}$ satisfying $AX = \mathbb{I}_p$ be given. Assume $y \in \mathbb{R}^p$ satisfies $Xy = 0$. Then, applying A yields

$$AXy = A0 = 0,$$

and because $AX = \mathbb{I}_p$, we find

$$0 = AXy = \mathbb{I}_p y = y.$$

So, $y = 0$ and we have shown that this is the only solution of $Xy = 0$. Hence, by the Invertible Matrix Theorem, X is nonsingular, and there is $Y \in \mathbb{R}^{p\times p}$ s.t. $XY = YX = \mathbb{I}_p$. Finally,

$$Y = \mathbb{I}_p Y = AXY = A\mathbb{I}_p = A.$$

Therefore, $XA = XY = \mathbb{I}_p$. □

Next, we'll recall some facts about the rank of matrices and use these to prove a few helpful results.

Lemma 2.4. For any $A \in \mathbb{R}^{p\times q}$ and $b \in \mathbb{R}^p$, we have

$$\text{rank}\,([A|b]) \geq \text{rank}(A).$$

Proof. Let the $p \times (q+1)$ matrix $[A'|b']$ be $\text{rref}\,([A|b])$ and recall that $\text{rank}(A)$ is equal to the number of linearly independent rows of A. Because rref is unique, it follows that $A' = \text{rref}(A)$. Furthermore, we know

$$\begin{aligned} \text{rank}(A') &= \text{the number of nonzero rows of } A' \\ \text{rank}([A'|b']) &= \text{the number of nonzero rows of } [A'|b']. \end{aligned}$$

Because A' is a submatrix of $[A'|b']$, if the kth row of $[A'|b']$ is exactly zero then the kth row of A' must also be exactly zero, and this holds for any $k = 1, ..., p$. Hence, A' cannot posses more nonzero rows than $[A'|b']$. Therefore, $\text{rank}([A'|b']) \geq \text{rank}(A')$, and since rref preserves rank it follows that

$$\text{rank}(A) = \text{rank}(A') \quad \text{and} \quad \text{rank}([A|b]) = \text{rank}([A'|b']).$$

Thus, we find

$$\text{rank}([A|b]) = \text{rank}([A'|b']) \ \geq \text{rank}(A') = \text{rank}(A).$$

□

Theorem 2.5 (Rank-Solvability)**.** Let $A \in \mathbb{R}^{p\times q}$ and $b \in \mathbb{R}^p$ be given. Then, exactly one of the following must hold:

(i) $\text{rank}([A|b]) > \text{rank}(A)$ and $Ax = b$ has no solution.

(ii) $\text{rank}([A|b]) = \text{rank}(A) = q$ and there exists a unique $x \in \mathbb{R}^q$ s.t. $Ax = b$.

(iii) $\text{rank}([A|b]) = \text{rank}(A) < q$ and there are infinitely many solutions of $Ax = b$.

Proof. By Lemma 2.4, $\text{rank}([A|b]) \geq \text{rank}(A)$ and hence either $\text{rank}([A|b]) > \text{rank}(A)$ or $\text{rank}([A|b]) = \text{rank}(A)$. In the latter case, either $\text{rank}(A) = q$ or $\text{rank}(A) < q$ since $A \in \mathbb{R}^{p\times q}$ can possess at most q linearly independent columns. In fact, because the rank of a matrix cannot exceed the number of its rows or columns, it follows that

$$\text{rank}(A) \leq \min\{p, q\}.$$

Regardless, we have three distinct cases to discuss.

Case 1: Assume $\text{rank}([A|b]) > \text{rank}(A)$.

Letting A' =rref(A) and $[A'|b']$ =rref$([A|b])$, we must have $\text{rank}([A'|b']) > \text{rank}(A')$ because rref preserves rank. Therefore, the number of nonzero rows of $[A'|b']$ must be strictly greater than the number of nonzero rows of A'. Thus, there exists $k \in \{1, ..., p\}$ such that the kth row of A' is exactly zero and the kth row of $[A'|b']$ is nonzero. This means the first q entries of the kth row of $[A'|b']$ must all be zero and the $(q+1)$st entry must be nonzero, i.e. we have a row represented as

$$[0 \ \cdots \ 0 \mid c \,]$$

with $c \neq 0$. This implies that the system has no solution as the resulting equation arising from this row is merely $0 = c$ with $c \neq 0$.

Case 2: Assume $\text{rank}([A|b]) = \text{rank}(A) = q$.

Then, we see that A has exactly q columns which form a linearly independent set. Let a_k represent the kth column of A for all $k = 1, ..., q$ so that $\{a_1, ..., a_q\}$ is linearly independent. Then, as $[A|b]$ consists of $(q+1)$ column vectors and $\text{rank}([A|b]) < q+1$, the set $\{a_1, ..., a_q, b\}$ must be linearly dependent. So, there are $y_1, ..., y_{q+1} \in \mathbb{R}$ not all equal to zero such that

$$\sum_{k=1}^{q} y_k a_k + y_{q+1} b = 0.$$

Now, if $y_{q+1} = 0$, then we find $\sum_{k=1}^{q} y_k a_k = 0$, which due to the linear independence of $\{a_1, ..., a_q\}$ implies $y_k = 0$ for every $k = 1, ..., q$. With this, it follows that the set $\{a_1, ..., a_q, b\}$ is linearly independent, contradicting our original assumption. Therefore, we conclude $y_{q+1} \neq 0$.

After some minor algebra, b can be expressed as a linear combination of vectors from the set $\{a_1, ..., a_q\}$ and letting $x_k = -\frac{y_k}{y_{q+1}}$ for every $k = 1, ..., q$, we find

$$\sum_{j=1}^{q} x_j a_j = b.$$

Expressed another way, this is just $Ax = b$ where $x = \begin{bmatrix} x_1 \\ \vdots \\ x_q \end{bmatrix}$.

Finally, the uniqueness of the solution x can be shown from a standard argument as follows. Let another solution $y \in \mathbb{R}^q$ satisfy $Ay = b$. Then, define $z = x - y$, and note that z satisfies

$$Az = A(x - y) = Ax - Ay = b - b = 0.$$

Or, stated another way,

$$\sum_{j=1}^{q} z_j a_j = 0.$$

Because A possesses q linearly independent columns, this equality implies $z_j = 0$ for all $j = 1, ..., q$. Thus, $z = 0$ and $y = x$ thereby yielding the uniqueness of the solution $x \in \mathbb{R}^q$.

Case 3: Assume $\text{rank}([A|b]) = \text{rank}(A) < q$.

The existence of a solution (not uniqueness) follows from a similar argument as Case 2 since $\text{rank}([A|b]) = \text{rank}(A)$ still holds. It remains to prove that the solution set is infinite.

We first prove that there are at least two solutions. Let $x \in \mathbb{R}^q$ satisfy $Ax = b$. Since $\text{rank}(A) < q$, we see that its columns, and thus the set $\{a_1, ..., a_q\}$, must be linearly dependent. Hence, there exist $y_1, .., y_q \in \mathbb{R}$ not all equal to zero such that

$$\sum_{j=1}^{q} y_j a_j = 0.$$

As before, this can be expressed as

$$Ay = 0$$

by constructing the vector $y \in \mathbb{R}^q$ from the scalar entries $y_1, ..., y_q$. In particular, since the entries of y cannot all be zero, we see that $y \neq 0$. With this, we find

$$A(x + y) = Ax + Ay = b + 0 = b.$$

So, $x + y$ also satisfies the linear system, and since $y \neq 0$, the solution $x + y$ is distinct from x. Therefore, we have at least two solutions.

Now that we are guaranteed the existence of two distinct solutions, we can further construct an infinite family of vectors that solve the linear system $Ax = b$. Let $x, y \in \mathbb{R}^q$ satisfy $Ax = b$ and $Ay = b$ with $x \neq y$. Consider, $x + kz$ where $z = x - y \neq 0$, and $k \in \mathbb{Z}$. Then, we find

$$\begin{aligned} A(x + kz) &= Ax + kAz \\ &= b + kA(x - y) \\ &= b + k(Ax - Ay) \\ &= b + k(b - b) \\ &= b - 0 \\ &= b. \end{aligned}$$

Thus, the set $\{x + kz : k \in \mathbb{Z}\}$ is an infinite family of solutions. □

Lemma 2.6. Let $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{q\times r}$ be given. Then,

$$\text{rank}(AB) \leq \min\{\text{rank}(A), \text{rank}(B)\}.$$

We postpone the proof of this theorem until later, but the result will be needed for the following result.

Theorem 2.7. Let $A \in \mathbb{R}^{p\times q}$ be given. Then, A^TA is nonsingular if and only if $\text{rank}(A) = q$.

Proof. As this is similar to a homework problem, we will prove only the forward direction of the theorem, and this will be done by contradiction. First, assume A^TA is nonsingular, but $\text{rank}(A) < q$. Then, by Lemma 2.6

$$\text{rank}(A^TA) \leq \text{rank}(A^T) = \text{rank}(A) < q.$$

However, $A^TA \in \mathbb{R}^{q\times q}$ is square, and thus by the IMT, $\text{rank}(A^TA) < q$ implies that A^TA is singular, contradicting our original assumption. Thus, $\text{rank}(A) = q$ if A^TA is nonsingular. □

Exercises - Review of Linear Algebra

Problem 2.1. Recall that for $A \in \mathbb{R}^{p\times p}$, the **trace** of A is defined by

$$\mathrm{tr}(A) = \sum_{k=1}^{p} A_{kk}.$$

Prove that if $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{q\times p}$, then

$$\mathrm{tr}(AB) = \mathrm{tr}(BA).$$

Problem 2.2. A square matrix A is called **skew-symmetric** if $A^T = -A$.

(a) Show that if A is any square matrix, then $A + A^T$ is symmetric and $A - A^T$ is skew-symmetric.

(b) Let A be a square matrix satifying $A = A_1 + A_2$ where A_1 is symmetric and A_2 is skew-symmetric. Find representations for both A_1 and A_2 in terms of A and A^T.

Problem 2.3. Let $A \in \mathbb{C}^{p\times p}$ and $B \in \mathbb{C}^{p\times p}$ be Hermitian matrices. Prove that AB is Hermitian if and only if A and B commute (i.e. $AB = BA$).

Problem 2.4. Let A, B, and $A+B$ be nonsingular matrices. Show that $A^{-1}+B^{-1}$ is nonsingular, as well. *Hint:* Compute a formula for $(A^{-1} + B^{-1})^{-1}$ first.

Problem 2.5. Let $A \in \mathbb{R}^{p\times q}$ and $B \in \mathbb{R}^{q\times r}$ be given. Prove

$$\mathrm{rank}(AB) \leq \mathrm{rank}(A).$$

Hint: Show $\mathrm{Col}(AB)$ is a subspace of $\mathrm{Col}(A)$.

Problem 2.6. Assume $A \in \mathbb{C}^{p\times q}$ satisfies $\mathrm{rank}(A) = q$. Show that $A^H A$ is nonsingular.

Problem 2.7. Let $A \in \mathbb{R}^{p\times q}$ and $b \in \mathbb{R}^p$ be given. Assume that $x_0 \in \mathbb{R}^q$ satisfies $Ax_0 = b$. Prove directly (i.e., without using the Invertible Matrix Theorem) that $y \in \mathbb{R}^q$ satisfies $Ay = b$ if and only if $y = x_0 + h$ for some $h \in \mathbb{R}^q$ satisfying $Ah = 0$.

Problem 2.8. Let $A \in \mathbb{R}^{p\times p}$ and $U, V \in \mathbb{R}^{p\times q}$ be given. Assume that A and the matrix $T = \mathbb{I} + V^T A^{-1} U$ are both nonsingular. Show that $A + UV^T$ is nonsingular with

$$(A + UV^T)^{-1} = A^{-1} - A^{-1}UT^{-1}V^T A^{-1}.$$

Chapter 3

Application: PageRank Algorithm

3.1 Introduction

PageRank is an algorithm proposed by Sergei Brin and Larry Page, the co-founders of Google, in the late 1990s. At its core, the PageRank algorithm serves as the basis for their famous Google search engine. The search algorithms used by Google are complex, and use several methods to return web pages based on search queries, but PageRank is one crucial portion of the overall structure. The purpose of this chapter is to describe the basic PageRank algorithm and its connection to Markov chains, and thus, Linear Algebra. The example in Brin and Page's paper [3] is the following. Suppose that we have a set of webpages connected by links. Now assume that each webpage has a descriptive title such as "Blaster the Burro". Suppose that a user is searching for the term "burro". There are likely to be a few pages that have this word in the title and the problem is to order these pages by relevance to the searcher's query. The purpose of PageRank, then, is to create a ranked list of websites in descending order of their connectivity so that we can identify which ones we expect to have the most web traffic, and hence which ones are most likely to be the website the user wants when they enter their search phrase.

One naive method – and, in fact, one used by many search engines before Google – was to simply count the number of occurrences of the word "burro" in each page and then sort the list of pages in descending order. This frequency based approach is easy to implement, but it does not reflect the network structure or connectivity of the internet. Additionally, it's easy for sites which rely on page hits for revenue to take advantage of such a method.

The theoretical formulation of PageRank is quite different from the frequency based approach. This algorithm assigns to each webpage w a number $r(w)$, representing a rank, such that $r(w) \geq 0$. Rescaling the ranks of all pages by a fixed positive constant does not affect their relative pageranks. So, these numbers are normalized such that $\sum_{w \in W} r(w) = 1$, where W denotes the set of all webpages. Therefore, $r(w)$ will be a probability distribution over all webpages.

In the next section, we will begin with the graph model of PageRank. We'll describe the notion of a directed graph and explain how PageRank can be stated in terms of the graph. Finally, we will show that the graph gives rise to a Markov chain

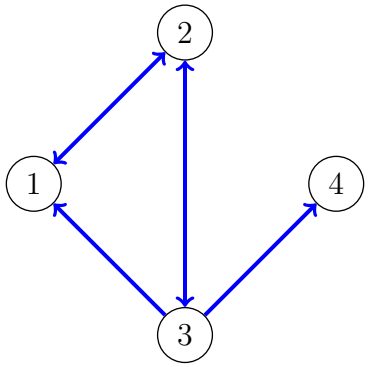
and that the PageRank is just the stationary distribution of the chain. Although PageRank is the motivating example, the notion of random walks on graphs and Markov chains have a wide variety of applications within mathematics, engineering, and the applied sciences.

3.2 PageRank Algorithm

3.2.1 Directed Graph Model

Suppose that w is a webpage and that we are browsing the web, currently focused on w. While viewing the page we may choose to click on a link and go to another page. Let F_w denote the set of forward links, i.e., the set of pages on the web to which w links. Also, let us denote by B_w the set of all backward links, i.e. pages that link to w or pages from which we can reach w. Given that we are at w we can randomly select one of the pages in F_w. Let N_w denote the number of pages in F_w and note that if we assume a uniform distribution, the probability of moving to any page $v \in F_w$ is given by $1/N_w$.

The mathematical abstraction into which this model fits is a random walk on a directed graph. A *directed graph* $\mathcal{G}$ is a pair of sets (V, E). The elements of V are called vertices. The set $E \subseteq V \times V$ is called the edge set. It often helps to draw a picture of a graph to better understand V and E. For instance suppose that $V = \{1, 2, 3, 4\}$ and that $E = \{(1, 2), (2, 1), (2, 3), (3, 2), (3, 4), (3, 1)\}$. We can visualize this graph with the following picture:



Note that the graph is directed, i.e., the edge $(3, 4)$ is different from the edge $(4, 3)$.

There are two ways to think of PageRank. We first explain the unrefined definition due to Brin and Page. Let $G = (V, E)$ be the webpage graph, where the vertices V are the individual webpages and the edges E are the links. So, there is an edge from the vertex u to the vertex v if and only if there is a hyperlink from u to v. Let B_u be the set of pages that have links to u (back links), let F_u be the pages to which u links (forward links), and let N_u be the size of F_u, so that N_u represents the number of pages to which u links. Notice that $v \in F_u$ if and only if $u \in B_v$.

Initially, each page is assigned a PageRank $r(u)$ by the following equation

$$r(u) = \sum_{v \in B_u} \frac{r(v)}{N_v}. \tag{3.1}$$

Thus, the PageRank of a page u is the sum of the PageRanks of all webpages that link to u divided by their respective number of forward links. A simple interpretation of (3.1) is that each page distributes its PageRank equally amongst all pages to which it points. In order to make this a probability distribution we normalize so that $\sum_{v \in V} r(v) = 1$ and $r(v) \geq 0$.

We are immediately faced with two problems. First, this definition is likely recursive; namely it appears as though in order to compute the PageRank of a specific page we need to know the PageRank of all sites to which it is backlinked. Of course, given the complexity of the network structure inherent in the internet, one of these pages (or one of its backlinked pages) could very well be a forward link of u. Hence, we would need to know $r(u)$ in order to compute $r(u)$, since this quantity would appear on both sides of the equation. The second issue concerns pages v that may not possess forward links at all, which would cause $N_v = 0$.

Let us number the vertices from 1 to n and denote the PageRank of the kth page by r_k so that r is a vector with n entries. Let W be the $n \times n$ matrix whose entries are defined by

$$w_{i,j} = \begin{cases} 1/N_j & \text{if } i \in F_j \\ 0 & \text{otherwise.} \end{cases} \tag{3.2}$$

Then, for the ith page the equation in (3.1) can be rewritten as

$$r_i = \sum_{j \in B_i} \frac{r_j}{N_j} = \sum_{i \in F_j} \frac{r_j}{N_j} = \sum_{j=1}^{n} w_{i,j} r_j.$$

If we let r be the (column) vector with entries $r_1, \ldots, r_n$, then the above equation (equating the left side and right side) can be written exactly as $r = Wr$. This means that r is just an eigenvector of the matrix W corresponding to the eigenvalue $\lambda = 1$, which isn't too difficult to compute using MATLAB even for fairly large n. Thus, computing the PageRank is actually an operation with which we are quite familiar from Linear Algebra. Next, we consider a second interpretation of PageRank and discuss the equation finally put forth by Brin and Page.

3.2.2 Random Surfer Model

Imagine a web surfer visiting the kth page on the web. At this point the surfer has a choice of N_k pages to visit, and decides to choose one by chance. With probability $1/N_k$ he or she chooses one of the pages from F_k at random and moves to that page. Now, the surfer arrives at page ℓ and is faced with a choice of N_ℓ pages. Again, through chance, the surfer chooses a page from F_ℓ at random with probability $1/N_\ell$. Therefore, at each page j he or she chooses one of the pages from F_j with a certain probability $p_{i,j}$ given by

$$p_{i,j} = \begin{cases} \frac{1}{N_j} & \text{if } i \in F_j \\ 0 & \text{otherwise.} \end{cases}$$

Thus, $p_{i,j}$ represents the probability of moving from site j to site i.

Imagine now that the surfer continues this random movement over the web. Over the long run (as $t \to \infty$) he or she will spend a certain fraction of time at each page. This fraction is then the stationary distribution of the Markov Chain with a transition matrix whose entries are $p_{i,j}$. Since finding the stationary distribution is usually the same problem as determining an eigenvector, we see that the solution to the random surfer model is exactly the solution to equation (3.1).

3.2.3 Brin and Page's Refinement of the Model

From either interpretation, there is still a problem with this model. If there is a page to which other pages point but which has no links pointing out (forward links) then the surfer will become "trapped". Consider, for example, the set of vertices and edges (V, E) in the previously-depicted directed graph. Vertex 4 possesses no forward links, and the probability of moving from vertex 4 to any other vertex must be 0. To avoid situations of this nature, Brin and Page introduce a damping factor, which allows the surfer to jump to any randomly-selected page on the web. Instead of (3.1), they ultimately defined the PageRank r_i by the equation

$$r_i = \alpha \sum_{j \in B_i} \frac{r_j}{N_j} + (1-\alpha)s_i \tag{3.3}$$

where $s = [s_1, \ldots, s_n]$ is ANY probability vector, meaning $s_i \geq 0$ and $s_1 + \cdots + s_n = 1$. Assuming that the probability of jumping to any other page is equally likely amongst the n pages, we choose $s_i = \frac{1}{n}$ for all $i = 1, ..., n$. The introduction of s_i now allows a random surfer to jump from one webpage to any other with equal probability $\frac{1-\alpha}{n}$. Therefore, the transition probabilities now become

$$g_{i,j} = \begin{cases} \frac{\alpha}{N_j} + \frac{(1-\alpha)}{n} & \text{if } i \in F_j \\ \frac{1-\alpha}{n} & \text{otherwise} \end{cases} \tag{3.4}$$

The damping factor α has been chosen so that $0 < \alpha < 1$ and this ensures that the numbers $\{g_{i,j} : i = 1, \ldots, n\}$ form a probability distribution for every $j = 1, ..., n$. Equation (3.3) can be written in terms of the previous matrix W as

$$r = \alpha W r + (1-\alpha)s = \alpha W r + (1-\alpha)Sr = [\alpha W + (1-\alpha)S]r$$

where S is the $n \times n$ matrix whose entries are all $\frac{1}{n}$. Note that, in practice, the number of websites n is a lot larger than N_j, which is the number of websites linked to by page j. So, the contribution from $\frac{1}{n}$ is small relative to the pages in F_j.

In general, we can associate to any graph its adjacency matrix $A = [a_{i,j}]$, where

$$a_{i,j} = \begin{cases} 1 & \text{if } i \in F_j \\ 0 & \text{otherwise.} \end{cases}$$

With this notation $N_j = \sum_{i=1}^{n} a_{i,j}$ for every $j = 1, ..., n$, and $w_{i,j} = \frac{1}{N_j} a_{i,j}$ for all $i, j = 1, ..., n$. Then, $G = [g_{i,j}]$ is defined as the matrix whose entries are given by (3.4) or equivalently by $G = \alpha W + (1-\alpha)S$. Note that all the entries of the matrix

G are non-negative and that the sum of the entries in each column of G is 1. The values $g_{i,j}$ can, as in the random model, be interpreted as the probability of moving from page j to page i. Finally, r can be determined exactly as the eigenvector of G corresponding to the eigenvalue $\lambda = 1$.

3.3 Stochastic Matrices

A matrix P with non-negative entries such that for every $j = 1, ..., n$

$$\sum_{i=1}^{n} p_{i,j} = 1$$

is called **stochastic**. It is not difficult to prove, given $0 \leq \alpha \leq 1$, and stochastic matrices P_1, P_2, that the convex combination of these matrices $P = \alpha P_1 + (1-\alpha)P_2$ is again stochastic.

This is the modification we made to the random model in (3.3) above. We have, on the one hand, a model P_1 in which the probabilities are assigned only to pages in F_i. We then consider the random model in which we are equally likely to visit any page from any other page and call this P_2. The model finally chosen is to consider the *Google matrix* $G = \alpha P_1 + (1-\alpha)P_2$ defined by (3.4).

Under suitable circumstances (e.g., if P is an **irreducible** stochastic matrix) the equation $P\pi = \pi$ is guaranteed to possess a solution π such that $\pi_i \geq 0$ for all $i = 1, ..., n$ and $\pi_1 + \cdots + \pi_n = 1$. The proof of this fact requires a lot more effort (see the Perron-Frobenius Theorem in [14]). However, for the special case of the Google matrix G we can give an ad-hoc proof.

Theorem 3.1. Assume $G \in \mathbb{R}^{n\times n}$ satisfies (3.4) for some $\alpha \in (0, 1)$ and $W \in \mathbb{R}^{n\times n}$ defined by (3.2). Then, there exists $\pi \in \mathbb{R}^n$ such that

1. $G\pi = \pi$
2. $\pi_i \geq 0$ for every $i = 1, ..., n$
3. $\sum_{i=1}^{n} \pi_i = 1.$

Proof. We begin by pointing out that if P is a stochastic matrix, then P^n is a stochastic matrix for any $n \geq 1$. The PageRank equation $\pi = G\pi$ can be written as a vector equation using (3.3), namely

$$\pi = (1-\alpha)s + \alpha W\pi \tag{3.5}$$

where W is the stochastic matrix defined by (3.2). Now multiply this equation by αW to get

$$\alpha W\pi = (1-\alpha)\alpha Ws + \alpha^2 W^2\pi.$$

If we substitute for $\alpha W\pi$ in (3.5) we get the equation

$$\pi = (1-\alpha)s + (1-\alpha)\alpha Ws + \alpha^2 W^2\pi. \tag{3.6}$$

Now multiply (3.5) by $\alpha^2 W^2$ to get the equation

$$\alpha^2 W^2 \pi = (1-\alpha)\alpha^2 W^2 s + \alpha^3 W^3 \pi.$$

Once again substitute this equation, but into (3.6), to get

$$\pi = (1-\alpha)s + (1-\alpha)\alpha W s + (1-\alpha)\alpha^2 W^2 s + \alpha^3 W^3 \pi.$$

If we continue this process we arrive at

$$\pi = (1-\alpha)\sum_{k=0}^{n} \alpha^k W^k s + \alpha^{n+1} W^{n+1} \pi.$$

Since W^k is stochastic for every $k = 1, ..., n+1$, the largest possible entry in W^k is 1. Additionally, $0 < \alpha < 1$ so each entry in the sequence of vectors $\alpha^{n+1} W^{n+1} \pi$ is dominated by a geometric sequence and therefore converges to 0. By taking the limit as $n \to \infty$ above, we obtain the formula

$$\pi = (1-\alpha)\sum_{k=0}^{\infty} \alpha^k W^k s.$$

Since α and W are known, this formula uniquely defines $\pi \in \mathbb{R}^n$. Note also that the non-negativity of the entries of $W^k s$ and the fact that $0 < \alpha < 1$ further imply that the series converges (by the ratio test). The entries of W^k and s are non-negative, and so $\pi_i \geq 0$ for every $i = 1, ..., n$ as well.

Let us denote the (i, j) entry of the matrix W^k by $w_{i,j}^{(k)}$. Then, write π_i as

$$\pi_i = (1-\alpha)\sum_{k=0}^{\infty} \alpha^k \sum_{j=1}^{n} w_{i,j}^{(k)} s_j.$$

Next, we sum the entries of π. Since all the series in question are convergent and have non-negative terms, they are absolutely convergent and we can interchange the summations to arrive at

$$\begin{aligned}\sum_{i=1}^{n} \pi_i &= (1-\alpha)\sum_{i=1}^{n}\sum_{k=0}^{\infty} \alpha^k \sum_{j=1}^{n} w_{i,j}^{(k)} s_j \\ &= (1-\alpha)\sum_{k=0}^{\infty} \alpha^k \left(\sum_{j=1}^{n} s_j \left(\sum_{i=1}^{n} w_{i,j}^{(k)} \right) \right).\end{aligned}$$

Since the matrix W^k is stochastic $\sum_{j=1}^{n} w_{i,j}^{(k)} = 1$. Because the vector s was chosen to be a probability distribution we also get $\sum_{j=1}^{n} s_j = 1$. Therefore, the above sum reduces to

$$\sum_{j=1}^{n} \pi_j = (1-\alpha)\sum_{k=0}^{\infty} \alpha^k = (1-\alpha)\frac{1}{1-\alpha} = 1.$$

□

For additional information regarding stochastic matrices, Perron-Frobenius theory, or the PageRank Algorithm, see [2, 12, 14, 15].

3.4 Summary of PageRank

Given a collection of n vertices (representing webpages) with $n \in \mathbb{N}$ and corresponding directed edges (representing links between websites), we perform the following steps to implement the final PageRank algorithm:

1. Enumerate the sites as
$$V = \{1, 2, 3, ..., n\}$$
and define the corresponding edge set
$$E = \{(i, j) : 1 \leq i, j \leq n \text{ and } i \to j\}$$

2. For $k = 1, ..., n$ define
$$F_k = \{\ell : (k, \ell) \in E\}$$

3. Define the adjacency matrix
$$A_{ij} = \begin{cases} 1 & i \in F_j \\ 0 & \text{else} \end{cases}$$
and the degree of each vertex $j = 1, ..., n$,
$$N_j = \sum_{i=1}^{n} A_{ij}.$$

4. From this let $W \in \mathbb{R}^{n \times n}$ be defined by
$$W_{ij} = \begin{cases} \frac{1}{N_j} & \text{if } i \in F_j \\ 0 & \text{else} \end{cases}$$

5. Choose a damping factor $\alpha \in (0, 1)$ and let $S \in \mathbb{R}^{n \times n}$ be defined by
$$S_{ij} = \frac{1}{n}$$
for every $i, j = 1, ..., n$.

6. Finally, compute the unique eigenvector of the matrix
$$G = \alpha W + (1 - \alpha)S$$
corresponding to the eigenvalue $\lambda = 1$, i.e. find $r \in \mathbb{R}^n$ s.t. $Gr = r$.
Then, rescale r such that
   (a) $r_i \geq 0$ for all $i = 1, ..., n$
   (b) $\sum_{i=1}^{n} r_i = 1$.

   The entries of r are exactly the respective PageRanks of the sites.

Exercises - PageRank

For problems which require computational simulation, please print and submit both your code and results (e.g., pictures).

Problem 3.1. Consider a graph with vertices

$$V = \{1, 2, 3, 4, 5, 6\}$$

and edges

$$\begin{aligned} E \;=\; & \{1 \to 2, 1 \to 3, 1 \to 4, 2 \to 1, 2 \to 3, 3 \to 4, \\ & 4 \to 1, 4 \to 3, 4 \to 5, 4 \to 6, 5 \to 6, 6 \to 5\}. \end{aligned}$$

(a) Draw the directed graph for this vertex set V and edge set E.

(b) Determine the adjacency matrix of the graph.

(c) The quantity $N_j = \sum_{i=1}^{n} a_{i,j}$ is called the degree of the vertex j. Compute the degree of each vertex of the graph.

(d) From this, compute the matrix W whose entries are given by

$$w_{i,j} = \begin{cases} 1/N_j & \text{if } i \in F_j \\ 0 & \text{otherwise.} \end{cases}$$

Then, use MATLAB to solve the equation $W\pi = \pi$ where $\pi_i \geq 0$ and $\sum_{i=1}^{6} \pi_i = 1$. Next, compute W^2, W^3, W^{10}, W^{30}. What do you observe?

(e) Now let $\alpha = 0.85$ and let $G = \alpha W + (1 - \alpha)S$, where S is the matrix with all of its entries equal to $1/6$. Solve the equation $G\pi = \pi$, where $\pi_i \geq 0$ and $\sum_{i=1}^{6} \pi_i = 1$. Compute G^2, G^3, G^{10}, G^{30}. What do you observe?

(f) Delete the vertices 5 and 6, as well as, any related edges and compute the resulting steady-state vector π for the associated Google matrix G.

Problem 3.2. Assume that $P, Q \in \mathbb{R}^{p \times p}$ are stochastic matrices.

(a) Let $0 \leq \alpha \leq 1$ be given, and prove that $R = \alpha P + (1 - \alpha)Q$ is a stochastic matrix.

(b) Prove that P^n is a stochastic matrix for every $n \in \mathbb{N}$.
Hint: Prove that the product of stochastic matrices is stochastic and then use induction.

Chapter 4

Linear Vector Spaces

4.1 Introduction and Definitions

Throughout we will take the field of scalars

$$\mathbb{K} = \mathbb{R} \text{ or } \mathbb{K} = \mathbb{C},$$

though other fields (e.g., finite fields) can also be used to define a vector space.

Definition 4.1. Assume $\mathcal{V}$ is a nonempty set (i.e. $\mathcal{V} \neq \emptyset$). Then, we say $\mathcal{V}$ is a **vector space** over $\mathbb{K}$ if:

1. There is an addition operation on $\mathcal{V}$, denoted by $+$, such that for all $u, v, w \in \mathcal{V}$
   (a) $u + v \in \mathcal{V}$
   (b) $u + v = v + u$
   (c) $(u + v) + w = u + (v + w)$.
2. There is a zero element, denoted $0 \in \mathcal{V}$, that satisfies
   $$u + 0 = u$$
   for every $u \in \mathcal{V}$.
3. For every $u \in \mathcal{V}$ there is an additive inverse element, $-u \in \mathcal{V}$, satisfying
   $$u + (-u) = 0.$$
4. There is scalar multiplication on $\mathbb{K}$ and $\mathcal{V}$ satisfying for all $u, v \in \mathcal{V}$ and $\alpha, \beta \in \mathbb{K}$
   (a) $\alpha u \in \mathcal{V}$
   (b) $\alpha(\beta u) = (\alpha\beta)u$
   (c) $\alpha(u + v) = \alpha u + \alpha v$
   (d) $(\alpha + \beta)u = \alpha u + \beta u$
   (e) $1 \cdot u = u$.

If $\mathbb{K} = \mathbb{R}$ then $\mathcal{V}$ is a **real vector space.** If $\mathbb{K} = \mathbb{C}$ then $\mathcal{V}$ is a **complex vector space.** The elements of $\mathcal{V}$ are called **vectors**. However, these are not to be confused with "traditional vectors" in $\mathbb{R}^n$ or $\mathbb{C}^n$ since the elements of a vector space can be functions, matrices, sequences, or any number of other objects, including the traditional vectors from Linear Algebra. Throughout, we will denote general vector spaces by $\mathcal{V}$ and $\mathcal{W}$.

Theorem 4.1. For any vector space $\mathcal{V}$,

1. 0 is unique in $\mathcal{V}$
2. For any $u \in \mathcal{V}$, the vector $-u$ is unique in $\mathcal{V}$.

Said another way, no vector space can possess two distinct additive identity vectors, nor can any vector possess two distinct additive inverses.

Proof. Assume that in addition to $0 \in \mathcal{V}$, there is an element $v \in \mathcal{V}$ such that

$$u + v = u \tag{4.1}$$

for every $u \in \mathcal{V}$. Then, we have

$$v = v + 0 = 0 + v = 0$$

where the first two equalities follow by the definition of a vector space, and the last equality follows from choosing $u = 0$ in (4.1). Hence, any element of $\mathcal{V}$ satisfying this property must be 0, which implies that the 0 element is unique.

Similarly, for any $u \in \mathcal{V}$

$$u + (-u) = 0;$$

hence, the uniqueness of $-u$ follows from the uniqueness of u and 0. □

Example 2 (Prominent Examples of Vector Spaces)**.** There are a number of familiar examples of linear vector spaces:

1. For $n \in \mathbb{N}$, $\mathbb{R}^n$ is a vector space over $\mathbb{R}$.
2. For $n \in \mathbb{N}$, $\mathbb{C}^n$ is a vector space over $\mathbb{R}$ or $\mathbb{C}$.

   We will see later that certain properties (e.g., dimension) of this space are dependent upon the scalars over which it is defined.
3. $F(\mathbb{R}) = \{\text{all functions } f : \mathbb{R} \to \mathbb{R}\}$ is a (infinite dimensional) vector space defined over $\mathbb{R}$.
4. $C(\mathbb{R}) = \{\text{all functions } f : \mathbb{R} \to \mathbb{R} \text{ which are continuous}\}$ is a vector space.
5. $C^\infty(\mathbb{R}) = \{\text{all functions } f : \mathbb{R} \to \mathbb{R} \text{ with continuous derivatives of any order}\}$ is a vector space.
6. $\mathbb{P}(\mathbb{R}) = \{\text{all polynomials } \ f : \mathbb{R} \to \mathbb{R}\}$ is a vector space.
7. For $n \in \mathbb{N}$, $\mathbb{P}^n(\mathbb{R}) = \{\text{all polynomials of degree} \leq n\}$ is a vector space.

8. For $p, q \in \mathbb{N}$, the set of all $p \times q$ matrices, $\mathbb{R}^{p\times q}$, is a vector space.

Definition 4.2. Let $\mathcal{V}$ be a vector space and $M \subseteq \mathcal{V}$. Assume that addition and scalar multiplication on M are defined by these same operations on $\mathcal{V}$. Then, M is a **subspace** of $\mathcal{V}$ if

1. $0 \in M$
2. M is closed under addition, i.e. for every $u, v \in M$, we have $u + v \in M$
3. M is closed under scalar multiplication, i.e. for all $\alpha \in \mathbb{K}$ and $v \in M$, we have $\alpha v \in M$.

In this case, M is also a vector space with the same operations.

Example 3. For any vector space $\mathcal{V}$, let $M = \{0\}$ where $0 \in \mathcal{V}$ is the additive identity. Then, M is a subspace of $\mathcal{V}$, called the **trivial subspace**.

Example 4. For every $n \in \mathbb{N}$, $\mathbb{P}^n(\mathbb{R})$ is a subspace of $\mathbb{P}(\mathbb{R})$; $\mathbb{P}(\mathbb{R})$ is a subspace of $C^\infty(\mathbb{R})$; $C^\infty(\mathbb{R})$ is a subspace of $C(\mathbb{R})$; and $C(\mathbb{R})$ is a subspace of $F(\mathbb{R})$.

Example 5. Given a vector space $\mathcal{V}$, let $T : \mathcal{V} \to \mathcal{V}$ satisfy the condition

$$T(\alpha u + \beta v) = \alpha T(u) + \beta T(v). \tag{4.2}$$

for all $\alpha, \beta \in \mathbb{K}$ and $u, v \in \mathcal{V}$. Then, $\mathcal{V}_0 = \{v \in \mathcal{V} : T(v) = 0\}$ is a subspace of $\mathcal{V}$. As a side note, any mapping satisfying (4.2) is called a **linear mapping**, and we'll study these in greater detail in subsequent chapters.

To show that $\mathcal{V}_0$ is a subspace, we merely demonstrate the subspace properties.

1. Show $0 \in \mathcal{V}_0$:

   Choose $\alpha = \beta = 0$, then $T(0) = 0$. Thus, $0 \in \mathcal{V}_0$.

2. Show $\mathcal{V}_0$ is closed under addition:

   Let $u, v \in \mathcal{V}_0$. Then, $T(u) = T(v) = 0$. Thus,
   $T(u + v) = T(u) + T(v) = 0 + 0 = 0$ by choosing $\alpha = \beta = 1$. Hence,
   $u + v \in \mathcal{V}_0$. Therefore $\mathcal{V}_0$ is closed under addition.

3. Show $\mathcal{V}_0$ is closed under scalar multiplication:

   Let $u \in \mathcal{V}_0$, $\alpha \in \mathbb{R}$ be given. Then, $T(v) = 0$. So, $T(\alpha u) = \alpha T(u) = \alpha \cdot 0 = 0$ by choosing $\beta = 0$ Hence, $\alpha u \in \mathcal{V}_0$, and $\mathcal{V}_0$ is closed under scalar multiplication.

Comment. This can be generalized to $T : \mathcal{V} \to \mathcal{W}$ where $\mathcal{W}$ is a vector space differing from $\mathcal{V}$.

4.2 Fundamental Properties of Vector Spaces

Definition 4.3. Let $\mathcal{V}$ be a vector space and $n \in \mathbb{N}$. For $v_1, ..., v_n \in \mathcal{V}$ the **span** of these vectors (i.e., the set of all linear combinations (or LCs) of the vectors $v_1, ..., v_n$) is denoted and defined as

$$\text{span}\{v_1, ..., v_n\} = \left\{ \sum_{j=1}^{n} \alpha_j v_j \;\middle|\; \alpha_j \in \mathbb{K} \;\text{ for all }\; j = 1, ..., n \right\}.$$

Theorem 4.2. Let $\mathcal{V}$ be a vector space and $n \in \mathbb{N}$. Then, for any $v_1, ..., v_n \in \mathcal{V}$,

$$M = \text{span}\{v_1, ..., v_n\}$$

is a subspace of $\mathcal{V}$.

Proof. Choosing $\alpha_j = 0$ for every $j = 1, ..., n$ shows $0 \in M$. Next, we show that M is closed under addition. Let $u, w \in M$ be given Then, there are $\alpha_i, \beta_i \in \mathbb{K}$ for all $i = 1, ..., n$ such that

$$u = \sum_{i=1}^{n} \alpha_i v_i, \qquad w = \sum_{i=1}^{n} \beta_i v_i.$$

Hence, we find

$$u + w = \sum_{i=1}^{n} \alpha_i v_i + \sum_{i=1}^{n} \beta_i v_i = \sum_{i=1}^{n} (\alpha_i + \beta_i) v_i = \sum_{i=1}^{n} \gamma_i v_i$$

where $\gamma_i = \alpha_i + \beta_i \in \mathbb{K}$ for all $i = 1, ..., n$, and thus $u + w \in \text{span}\{v_1, ..., v_n\} = M$. Finally, to show that M is closed under scalar multiplication, we use the same type of argument so that for any given $u \in M$ and $k \in \mathbb{K}$, there is $\alpha_i \in \mathbb{K}$ for all $i = 1, ..., n$ such that

$$ku = k \sum_{i=1}^{n} \alpha_i v_i = \sum_{i=1}^{n} k\alpha_i v_i = \sum_{i=1}^{n} \gamma_i v_i$$

where $\gamma_i = k\alpha_i \in \mathbb{K}$ for all $i = 1, ..., n$, and thus $ku \in \text{span}\{v_1, ..., v_n\} = M$.

□

Definition 4.4. Let $\mathcal{V}$ be a vector space and $M \subseteq \mathcal{V}$ be a subspace of $\mathcal{V}$. Take $n \in \mathbb{N}$ and let $S = \{v_1, ..., v_n\} \subseteq M$. Then, we say that S **spans** M (or is a **spanning set** of M) if for every $v \in M$ there exists $\alpha_1, ..., \alpha_n \in \mathbb{K}$ such that

$$v = \sum_{j=1}^{n} \alpha_j v_j.$$

Example 6. Let $e_k \in \mathbb{R}^p$ for any $k = 1, 2, ..., p$ be defined by

$$(e_k)_i = \delta_{ik} := \begin{cases} 1, & \text{if } i = k \\ 0, & \text{else.} \end{cases}$$

Here, δ_{ik} is referred to as the Kronecker delta. Written another way, we define the vectors

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad \dots, \qquad e_p = \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

to be the unit vectors with 1 in the kth entry and 0 in all other entries. Then, for any $p \in \mathbb{N}$, $B = \{e_1, ..., e_p\}$ is a spanning set for $\mathbb{R}^p$.

Example 7. Let $\mathcal{V} = \mathbb{R}^2$ and define

$$S_1 = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\} \qquad S_2 = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}.$$

Of course, $M = \text{span}(S_1) = \mathcal{V}$ is a subspace and both S_1 and S_2 span M.

Example 8. Let $\mathcal{V} = \mathbb{R}^3$ and define

$$M = \left\{ v = \begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix} : v_1, v_2 \in \mathbb{R} \right\}.$$

Then, by definition M is a subspace of $\mathcal{V}$. Can we find a spanning set of M? Clearly, we can decompose any element $v \in M$ into

$$v = \begin{bmatrix} v_1 \\ v_2 \\ 0 \end{bmatrix} = v_1 \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + v_2 \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} = v_1 e_1 + v_2 e_2$$

where $v_1, v_2 \in \mathbb{R}$ and $e_1, e_2 \in \mathbb{R}^3$ are defined above. Thus, $S = \{e_1, e_2\}$ spans M.

Definition 4.5. Let $\mathcal{V}$ be a vector space. Take $n \in \mathbb{N}$ and let $S = \{v_1, ..., v_n\} \subseteq \mathcal{V}$.

1. S is **linearly independent** (abbreviated **LI**) if

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \quad \text{implies} \quad \alpha_j = 0 \text{ for all } j = 1, ..., n.$$

2. S is **linearly dependent** (abbreviated **LD**) if S is not linearly independent; that is, if there are $\alpha_i \in \mathbb{K}$ for $i = 1, ..., n$ that are not ALL zero such that

$$\sum_{j=1}^{n} \alpha_j v_j = 0.$$

Comment. For $S \subseteq \mathbb{C}^p$, we have good machinery for determining whether or not a set of vectors is linearly independent. Outside of this setting, things become tricky.

Example 9. Let $\mathcal{V} = \mathbb{R}^3$ and consider $S = \{v_1, v_2, v_3\}$ where

$$v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \quad v_3 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}.$$

Is S linearly dependent or independent?
Obviously, we consider $\sum_{j=1}^{n} \alpha_j v_j = 0$, which is equivalent to the linear system

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \alpha_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

We perform Gaussian elimination on the matrix to reduce it to rref and find

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 1 & 1 \end{bmatrix} \sim \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

and because the resulting system possesses infinitely many solutions, we see that not all of the α coefficients need be zero to satisfy the system of equations. Hence, S is linearly dependent.

Of course, if we make a small change to v_2 and instead consider the set $T = \{v_1, e_2, v_3\}$, then rref of the resulting matrix is instead

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

and we find a unique solution, namely the zero solution. Hence, T is linearly independent.

Theorem 4.3. Let $\mathcal{V}$ be a vector space, $n \in \mathbb{N}$ and $S = \{v_1, ..., v_n\} \subseteq \mathcal{V}$.

1. If $n = 1$, then $S = \{v_1\}$ is linearly independent if and only if $v_1 \neq 0$
2. If $n \geq 2$, then S is linearly dependent if and only if <u>at least</u> one of the vectors is a linear combination of the remaining vectors in S, i.e. there is $k \in \{1, ..., n\}$ such that
$$v_k = \sum_{\substack{j=1 \\ j \neq k}}^{n} \alpha_j v_j$$
for some $\alpha_j \in \mathbb{K}$, $j = 1, ..., n$.
3. If $0 \in S$, then S is linearly dependent.
4. If S is linearly independent, then any nonempty subset of S is also linearly independent.
5. If S is linearly dependent, $m \in \mathbb{N}$ with $m > n$, and $v_{n+1}, ..., v_m \in \mathcal{V}$, then $\{v_1, ..., v_m\}$ is also linearly dependent.

The fourth conclusion displays that any subset of a linearly independent set must also be linearly independent, while the fifth conclusion shows that any extension of a linearly dependent set is also linearly dependent.

Proof. The first and third implications are straightforward. We prove the second conclusion, and the final two conclusions will be assigned as homework exercises (cf. Problem 4.5).

To prove the backward direction, first assume

$$v_k = \sum_{\substack{j=1 \\ j \neq k}}^{n} \alpha_j v_j$$

for some k and $\alpha_j \in \mathbb{K}$ with $j \neq k$. Then, subtracting v_k to the right side yields a linear combination of the v_j vectors that is equal to zero. Hence, if we define new coefficients

$$\beta_j = \begin{cases} \alpha_j, & j \neq k \\ -1, & j = k, \end{cases}$$

we see that $\sum_{j=1}^{n} \beta_j v_j = 0$. However, it is not the case that $\beta_j = 0$ for all $j = 1, ..., n$. Therefore, S is not linearly independent, and so must be linearly dependent.

Now, to prove the forward implication, assume S is linearly dependent. Then, there exists $\alpha_j \in \mathbb{K}$ for all $j = 1, ..., n$ such that $\sum_{j=1}^{n} \alpha_j v_j = 0$ and not all of the α_j's are zero. Since there is a nonzero α_j, let's denote its index by k, so that $\alpha_k \neq 0$. Thus, we rewrite the sum as

$$\sum_{\substack{j=1 \\ j \neq k}}^{n} \alpha_j v_j + \alpha_k v_k = 0.$$

or by subtracting the v_k term and dividing by $\alpha_k \neq 0$, equivalently

$$\begin{aligned} v_k &= -\frac{1}{\alpha_k} \sum_{\substack{j=1 \\ j \neq k}}^{n} \alpha_j v_j \\ &= \sum_{\substack{j=1 \\ j \neq k}}^{n} \left(-\frac{\alpha_j}{\alpha_k} \right) v_j \\ &= \sum_{\substack{j=1 \\ j \neq k}}^{n} \beta_j v_j \end{aligned}$$

where $\beta_j = -\frac{\alpha_j}{\alpha_k}$ for all $j = 1, ..., n$ with $j \neq k$. Therefore, at least one of $v_1, ..., v_n$ can be written as a linear combination of the remaining vectors in S. □

Definition 4.6. A **basis** for a vector space $\mathcal{V}$ is a linearly independent subset of $\mathcal{V}$ that also spans $\mathcal{V}$; that is, a set $B = \{v_1, ..., v_n\} \subseteq \mathcal{V}$ is a basis if and only if

1. For any $u \in \mathcal{V}$, there are $\alpha_1, ..., \alpha_n \in \mathbb{K}$ such that $u = \sum_{j=1}^{n} \alpha_j v_j$

2. If $\sum_{j=1}^{n} \alpha_j v_j = 0$ for some $\alpha_1, ..., \alpha_n \in \mathbb{K}$ then $\alpha_j = 0$ for every $j = 1, ..., n$.

Example 10. Recall $e_k \in \mathbb{R}^p$ for any $k = 1, 2, ..., p$ defined by

$$(e_k)_i = \begin{cases} 1, & \text{if } i = k \\ 0, & \text{else.} \end{cases}$$

Then, for any $p \in \mathbb{N}$, $B = \{e_1, ..., e_p\}$ is a basis for $\mathbb{R}^p$. In fact, B is generally referred to as the standard basis for $\mathbb{R}^p$

Example 11. Let $A = [1, 1, 1] \in \mathbb{R}^{1,3}$ and define $\mathcal{V}_0 = \{x \in \mathbb{R}^3 : Ax = 0\}$ Can we find a basis for $\mathcal{V}_0$?

First, consider $x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} \in \mathcal{V}_0$, then

$$Ax = \begin{bmatrix} 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

is equivalent to

$$x_1 + x_2 + x_3 = 0$$

and therefore

$$x_1 = -x_2 - x_3.$$

This leads to the decomposition

$$x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -x_2 - x_3 \\ x_2 \\ x_3 \end{bmatrix} = x_2 \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + x_3 \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}.$$

Since $x \in \mathcal{V}_0$ was arbitrary we can write any element this way, for any constants $x_2, x_3 \in \mathbb{R}$, and thus

$$\mathcal{V}_0 = \left\{ s \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + t \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} : s, t \in \mathbb{R} \right\}$$

So, now we consider the set of vectors

$$S = \left\{ \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix} \right\}.$$

We know that S spans $\mathcal{V}_0$, which means we need only show that S is linearly independent. Let

$$v_1 = \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix}, \quad v_2 = \begin{bmatrix} -1 \\ 0 \\ 1 \end{bmatrix}$$

and consider $\alpha_1 v_1 + \alpha_2 v_2 = 0$, or to write this another way

$$\begin{bmatrix} -\alpha_1 - \alpha_2 \\ \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}.$$

Solving the system of equations, we find $\alpha_1 = \alpha_2 = 0$ and hence S is linearly independent. Since S is both a spanning set for $\mathcal{V}_0$ and linearly independent, it forms a basis for $\mathcal{V}_0$.

Theorem 4.4 (Spanning Set Theorem)**.** Let $\mathcal{V}$ be a vector space and $S = \{v_1, ..., v_p\}$ be a spanning set for a subspace $M \subseteq \mathcal{V}$. Then,

1. If for some $k = 1, .., p$ the element v_k can be written as a linear combination of the remaining vectors in S, then the set

$$S' = S \setminus \{v_k\}$$

also spans M.

2. If $M \neq \{0\}$ then there is $T \subseteq S$ such that T is a basis for M.

Proof. To prove the first conclusion, we fix k such that v_k is a linear combination of the remaining elements of S. Then, there are $\alpha_1, ..., \alpha_{k-1}, \alpha_{k+1}, ...\alpha_p \in \mathbb{K}$ such that

$$v_k = \sum_{\substack{j=1 \\ j \neq k}}^{p} \alpha_j v_j.$$

Hence, letting $u \in M$ be given, we find $\beta_1, ..., \beta_p \in \mathbb{K}$ such that

$$u = \sum_{j=1}^{p} \beta_j v_j$$

since S spans M. Of course, substituting the previous representation for v_k, we can rewrite this as

$$\begin{aligned} u &= \sum_{\substack{j=1 \\ j \neq k}}^{p} \beta_j v_j + \beta_k v_k \\ &= \sum_{\substack{j=1 \\ j \neq k}}^{p} \beta_j v_j + \beta_k \sum_{\substack{j=1 \\ j \neq k}}^{p} \alpha_j v_j \\ &= \sum_{\substack{j=1 \\ j \neq k}}^{p} \gamma_j v_j. \end{aligned}$$

where $\gamma_j = \beta_j + \beta_k \alpha_j$ for every $j = 1, .., p$ and $j \neq k$. Thus, u can be expressed as a linear combination of elements from the set $S' = S \setminus \{v_k\}$, and since $u \in M$ was arbitrary, we see that S' spans M.

Next, we prove the second conclusion, in part, by using this first result. Notice that if S is linearly independent, then it is already a basis for M and the result follows as $S \subseteq M$. Otherwise, S is a linearly dependent set, which provides two possible cases by Theorem 4.3 - either $S = \{0\}$ or S possesses at least two elements and at least one of these elements can be written as a linear combination of the remaining vectors in S. Certainly, the first possibility yields $M = \{0\}$, which is false by assumption. Therefore, the second scenario must be true, and using the first conclusion of this theorem, we may continue to remove elements of S until we have constructed a linearly independent subset of $S' \subseteq S$ that still spans M. As such, S' must be a basis for M, and the proof is complete.

□

From Theorems 4.3 and 4.4, we see that bases can be constructed in two distinct manners - from a top-down approach or bottom-up approach. In particular, we

may build a basis for a vector space or subspace by beginning with a linearly independent subset and adding new vectors that maintain the linear independence of the resulting set until we finally arrive at a spanning set, which must then be a basis. Alternatively, we may begin with a spanning set for the subspace - perhaps even beginning with every element of the vector space - and then remove elements of the spanning set while maintaining the spanning property of the resulting set until we arrive at one that is also linearly independent, and hence a basis.

One reason why we care about bases is that, unlike a mere spanning set, a basis can uniquely represent every element of the vector space or subspace that it spans.

Theorem 4.5. Let $\mathcal{V}$ be a vector space, $n \in \mathbb{N}$, and $B = \{v_1, ..., v_n\}$ be a basis for $\mathcal{V}$. Then, for any $v \in \mathcal{V}$, there are unique $\alpha_1, ..., \alpha_n \in \mathbb{K}$ such that $v = \sum_{j=1}^{n} \alpha_j v_j$.

Proof. The existence of $\alpha_1, ..., \alpha_n \in \mathbb{K}$ satisfying this property follows directly from B being a basis, and thus spanning $\mathcal{V}$. To prove uniqueness, let $\beta_1, ..., \beta_n \in \mathbb{K}$ also satisfy $v = \sum_{j=1}^{n} \beta_j v_j$. Subtracting these two different representations of v, we find

$$\sum_{j=1}^{n} \alpha_j v_j - \sum_{j=1}^{n} \beta_j v_j = v - v = 0.$$

However, defining $\gamma_j = \alpha_j - \beta_j$ leads to $\sum_{j=1}^{n} \gamma_j v_j = 0$, and since $B = \{v_1, ..., v_n\}$ is linearly independent, we must have $\gamma_j = 0$ for all $j = 1, ..., n$. With this, we have $\alpha_j = \beta_j$ for every $j = 1, ..., n$, and thus the representation is unique. □

Definition 4.7. The $\alpha_1, ..., \alpha_n \in \mathbb{K}$ guaranteed by Theorem 4.5 are called the **coordinates of $v \in \mathcal{V}$ with respect to the basis** B. Of course, a basis for a vector space $\mathcal{V}$ need not be unique and changing the basis will change these coordinates.

Example 12. Recall the vector space of quadratic polynomials

$$\mathbb{P}^2 = \left\{ f : \mathbb{R} \to \mathbb{R} \;\middle|\; f(x) = a_0 + a_1 x + a_2 x^2 \text{ for some } a_0, a_1, a_2 \in \mathbb{R} \right\}.$$

Then, the set $B_2 = \{1, x, x^2\}$ is a basis for $\mathbb{P}^2$ and the resulting coordinates of any quadratic polynomial with respect to B_2 are $\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix} \in \mathbb{R}^3$. Notice that the ordering of the basis matters to identify the associated coordinates. Analogously, the set $B_n = \{1, ..., x^n\}$ is a basis for $\mathbb{P}^n$ with corresponding coordinates $\begin{bmatrix} a_0 \\ \vdots \\ a_n \end{bmatrix} \in \mathbb{R}^{n+1}$.

Example 13. Recall the vector space of real-valued 2×2 matrices

$$\mathbb{R}^{2\times 2} = \left\{ \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \;\middle|\; a_{11}, a_{12}, a_{21}, a_{22} \in \mathbb{R} \right\}.$$

Then, the set

$$B = \left\{ \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \right\}$$

is a basis for $\mathbb{R}^{2\times 2}$ and the resulting coordinates of any matrix with respect to this basis are $\begin{bmatrix} a_{11} \\ a_{12} \\ a_{21} \\ a_{22} \end{bmatrix} \in \mathbb{R}^4$.

These two examples illustrate that the spaces $\mathbb{P}^2$ and $\mathbb{R}^3$, and the spaces $\mathbb{R}^{2\times 2}$ and $\mathbb{R}^4$, respectively, are quite alike since there is an explicit coordinate mapping between them. In fact, such a coordinate mapping is a specific type of linear transformation defined below.

Definition 4.8. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces and $T : \mathcal{V} \to \mathcal{W}$.

1. T is **one-to-one** if for any $u, v \in \mathcal{V}$

$$T(u) = T(v) \quad \text{implies} \quad u = v.$$

2. T is **onto** if for any $w \in \mathcal{W}$ there is $v \in \mathcal{V}$ such that $T(v) = w$.

Definition 4.9. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces. A mapping $T : \mathcal{V} \to \mathcal{W}$ is called an **isomorphism** if T is linear, one-to-one, and onto. In this case, we say that $\mathcal{V}$ and $\mathcal{W}$ are **isomorphic**.

Isomorphisms are important mappings as they preserve the fundamental algebraic properties of vector spaces. Hence, the two vector spaces $\mathcal{V}$ and $\mathcal{W}$ in the definition above behave in similar ways. Though we won't study the topic in much detail, (group) isomorphisms are fundamental objects within Abstract Algebra and even arise naturally in some analytic frameworks. For instance, the Laplace Transform, which many of us have used to solve linear differential equations, is an isomorphism between certain vector spaces of functions.

Returning to Example 12, we see that $\mathbb{P}^2$ and $\mathbb{R}^3$ are isomorphic. More generally, if a vector space has a basis B consisting of p elements, then we may naturally associate it with $\mathbb{K}^p$ (often $\mathbb{R}^p$) using the coordinate mapping generated by B.

Theorem 4.6. Let $B = \{v_1, ..., v_p\}$ be a basis for a vector space $\mathcal{V}$. Then, the coordinate mapping $T : \mathcal{V} \to \mathbb{K}^p$ defined for any $v \in \mathcal{V}$ by $T(v) = \alpha^v$, where $\alpha_1^v, ..., \alpha_p^v$ are the coordinates of v with respect to B and

$$\alpha^v = \begin{bmatrix} \alpha_1^v \\ \vdots \\ \alpha_p^v \end{bmatrix},$$

is an isomorphism.

Proof. By Theorem 4.5, for every $v \in \mathcal{V}$ there are unique $\alpha_1^v, ..., \alpha_p^v$ such that

$$v = \sum_{j=1}^{p} \alpha_j^v v_j.$$

Thus, to prove the theorem, we need to show that the function $T : \mathcal{V} \to \mathbb{K}^p$ defined by this coordinate mapping

$$T(v) = \begin{bmatrix} \alpha_1^v \\ \vdots \\ \alpha_p^v \end{bmatrix} =: \alpha^v,$$

is linear, one-to-one, and onto.

To show that T is linear, we merely let $u, w \in \mathcal{V}$ be given and find $\alpha^u, \alpha^w \in \mathbb{K}^p$ such that

$$u = \sum_{j=1}^{p} \alpha_j^u v_j \qquad \text{and} \qquad w = \sum_{j=1}^{p} \alpha_j^w v_j.$$

This immediately implies

$$u + w = \sum_{j=1}^{p} (\alpha_j^u + \alpha_j^w) v_j$$

and thus

$$T(u + w) = \alpha^u + \alpha^w = T(u) + T(w).$$

The scalar multiplication property follows analogously, and T is linear.

We note that the one-to-one and onto properties actually follow from the linearity of T and the linear properties of the vector space $\mathcal{V}$. Indeed, $T(u) = T(v)$ implies $\alpha^u = \alpha^v$, and since every coefficient of the linear combination is equal, we see that $u = v$. Additionally, every $\alpha \in \mathbb{K}^p$ must give rise to a vector $v \in \mathcal{V}$ as $v_j \in \mathcal{V}$ for every $j = 1, .., p$ and $\mathcal{V}$ is closed under both addition and scalar multiplication.

We refer the interested reader to Lay [11], pp. 219-220 for additional details. □

Theorem 4.7. Let $\mathcal{V}$ be a vector space and suppose that $\{v_1, ..., v_p\} \subseteq \mathcal{V}$ spans $\mathcal{V}$ while $\{u_1, ..., u_q\} \subseteq \mathcal{V}$ is linearly independent. Then, $q \leq p$.

Proof. First, if $\mathcal{V} = \{0\}$ then $v_1 = 0$ and the result holds trivially. Next, assume $\mathcal{V} \neq \{0\}$, then assume $q > p$ and work to derive a contradiction. Given the spanning set $\{v_1, ..., v_p\}$ we may use Theorem 4.4 to remove vectors from this set until arriving at a basis for $\mathcal{V}$ consisting of m elements (with $m \leq p$) and denoted by $B \subseteq \{v_1, ..., v_p\}$. Then, using Theorem 4.5, we may uniquely represent every element of the linearly independent set $\{u_1, ..., u_q\}$ as a linear combination of elements of B. For every $k = 1, ..., q$, the coefficients of this unique representation are exactly the coordinates of u_k with respect to the basis B. Hence, we now have a set of coordinates denoted $\{w_1, ..., w_q\} \subseteq \mathbb{K}^m$ that are generated using the coordinate mapping $T : \mathcal{V} \to \mathbb{K}^m$, which satisfies $T(u_k) = w_k$ for every $k = 1, ..., q$. Constructing a matrix $A \in \mathbb{K}^{m \times q}$ using the vector w_k as the kth column of A, we note that $\text{rank}(A) \leq m$ and thus

$$\text{rank}\left([A|0]\right) = \text{rank}(A) \leq m \leq p < q.$$

Thus, by the Rank-Solvability Theorem (Theorem 2.5), the system $Ax = 0$ has infinitely many solutions. From these, we choose a nonzero solution $\beta \in \mathbb{K}^q \setminus \{0\}$. Of course, this demonstrates that the set of coordinate vectors $\{w_1, ..., w_q\}$ is linearly dependent as $A\beta = 0$ is equivalent to $\sum_{j=1}^{q} \beta_j w_j = 0$ and we know $\beta \neq 0$. Finally, using Theorem 4.6 and the inverse coordinate mapping $T^{-1} : \mathbb{K}^m \to \mathcal{V}$ generated by the basis B, this demonstrates that the set $\{u_1, ..., u_q\}$ is linearly dependent, thereby providing a contradiction. Hence, $q \leq p$.

For further details, consult one of the references, such as Noble & Daniel [16] (Theorem 5.28, p. 197) or Lay [11] (Theorem 9 in Section 4.5, p. 225). □

Theorem 4.8. Let $\mathcal{V}$ be a vector space with a basis consisting of p vectors $\{v_1, ..., v_p\}$. Then, every basis of $\mathcal{V}$ must possess exactly p vectors.

Proof. Let $\{u_1, ..., u_q\}$ be any basis for $\mathcal{V}$. By Theorem 4.7, since $\{v_1, ..., v_p\}$ spans $\mathcal{V}$ and $\{u_1, ..., u_q\} \subseteq \mathcal{V}$ is linearly independent, we find $q \leq p$. By Theorem 4.7, since $\{v_1, ..., v_p\}$ is linearly independent and $\{u_1, ..., u_q\}$ spans $\mathcal{V}$ we find $p \leq q$. Finally, $p = q$ and since $\{u_1, ..., u_q\}$ was an arbitrary basis, we see that every basis must have exactly p elements. □

Definition 4.10. The **dimension** of a vector space $\mathcal{V}$ denoted $\dim(\mathcal{V})$ is defined to be the number of vectors in any basis. If $\mathcal{V}$ does not possess a basis with finitely many elements, then we say $\dim(\mathcal{V}) = \infty$.

The idea of dimension has a few useful implications, especially for subspaces.

Theorem 4.9. Let $\mathcal{V}$ be a vector space with $\dim(\mathcal{V}) = p \in \mathbb{N}$ and let $M \subseteq \mathcal{V}$ be a subspace. Then, $\dim(M) \leq p$.

Proof. Let $B = \{v_1, ..., v_p\}$ be a basis for $\mathcal{V}$ and $B_0 = \{u_1, ..., u_q\}$ be a basis for M so that $\dim(M) = q$. Since $B_0 \subseteq \mathcal{V}$ is a basis for M, it must be a linearly independent subset of $\mathcal{V}$. Additionally, since B is a basis for $\mathcal{V}$, it must also be a spanning set. By Theorem 4.7, it follows that $q \leq p$, and thus

$$\dim(M) = q \leq p = \dim(\mathcal{V}).$$

□

With this result, we may now easily prove Lemma 2.6, which was stated in the Review Exercises section.

Lemma (Lemma 2.6)**.** Let $A \in \mathbb{R}^{p \times q}$ and $B \in \mathbb{R}^{q \times r}$ be given. Then,

$$\operatorname{rank}(AB) \leq \min\{\operatorname{rank}(A), \operatorname{rank}(B)\}.$$

Proof. By definition, we have

$$\begin{aligned} \operatorname{rank}(AB) &= \dim(\operatorname{Col}(AB)) = \dim\left(\{ABx : x \in \mathbb{R}^r\}\right) \\ \operatorname{rank}(A) &= \dim(\operatorname{Col}(A)) = \dim\left(\{Ay : y \in \mathbb{R}^q\}\right). \end{aligned}$$

Notice first that $\operatorname{Col}(AB) \subseteq \operatorname{Col}(A)$. Indeed, if $z \in \operatorname{Col}(AB)$, then there exists $x \in \mathbb{R}^r$ such that $z = ABx$. Letting $y = Bx \in \mathbb{R}^q$, we can merely write $z = Ay$ and

thus $z \in \text{Col}(A)$. Next, since both $\text{Col}(AB)$ and $\text{Col}(A)$ are not merely subsets, but in fact, subspaces of $\mathbb{R}^p$, it follows that $\text{Col}(AB)$ is a subspace of $\text{Col}(A)$. By Theorem 4.9, we find

$$\dim(\text{Col}(AB)) \leq \dim(\text{Col}(A))$$

or stated another way

$$\text{rank}(AB) \leq \text{rank}(A). \tag{4.3}$$

It remains to show $\text{rank}(AB) \leq \text{rank}(B)$ in order to complete the proof, and we'll use a result from Linear Algebra, namely $\text{rank}(C^T) = \text{rank}(C)$ for any $C \in \mathbb{R}^{p\times q}$, along with (4.3) to do so. In particular, since (4.3) holds for any product of matrices, we may apply it to $B^T A^T$ to find

$$\text{rank}(B^T A^T) \leq \text{rank}(B^T).$$

This then yields

$$\text{rank}(AB) = \text{rank}((AB)^T) = \text{rank}(B^T A^T) \leq \text{rank}(B^T) = \text{rank}(B)$$

so that $\text{rank}(AB) \leq \text{rank}(B)$ and combining this upper bound with (4.3) completes the proof. □

Lemma 4.10. Let $\mathcal{V}$ be a vector space with two subspaces $\mathcal{V}_0 \subseteq \mathcal{V}_1 \subseteq \mathcal{V}$ satisfying $\dim(\mathcal{V}_0) = \dim(\mathcal{V}_1) < \infty$. Then, $\mathcal{V}_0 = \mathcal{V}_1$.

Proof. We assign this as a homework exercise (cf. Problem 4.7). Generally, the idea is that if there is an element u in $\mathcal{V}_1$ that is not in $\mathcal{V}_0$, then we could add u to the basis for $\mathcal{V}_0$ and create a linearly independent set within $\mathcal{V}_1$ containing more elements than the basis for $\mathcal{V}_1$, which provides a contradiction. □

As the next example demonstrates, the field over which a vector space is defined can be crucial to its inherent properties, including its dimension.

Example 14. Consider $\mathcal{V}_1 = \mathbb{C}^p$ defined over $\mathbb{K} = \mathbb{C}$. Then, $\dim(\mathbb{C}^p) = p$ because

$$S = \{e_k : k = 1, ..., p\}$$

is a basis for $\mathcal{V}_1$. Next, consider $\mathcal{V}_2 = \mathbb{C}^p$ defined over $\mathbb{K} = \mathbb{R}$. Then, $\dim(\mathbb{C}^p) = 2p$ because

$$S = \{e_k,\, ie_k : k = 1, ..., p\}$$

is a basis. Of course, the elements in the vector space are identical, but because the field of scalars over which they are defined differ, $\mathcal{V}_1$ and $\mathcal{V}_2$ are, in fact, different vector spaces, and in particular, have differing dimensions.

4.3 Infinite Dimensional Spaces

All of the bases that we've mentioned previously have finitely many elements. If we consider an infinite-dimensional vector space $\mathcal{V}$, can we still identify a basis B? What do the two properties (linear independence and spanning $\mathcal{V}$) look like if $\dim(\mathcal{V}) = \infty$? It turns out that at least one of these properties must be reformulated to account for the fact that B must contain infinitely many elements.

Definition 4.11. If $\dim(\mathcal{V}) = \infty$ then we say $B \subseteq \mathcal{V}$ is a **Hamel basis** for $\mathcal{V}$ if

1. Every finite subset of B satisfies the linear independence property; namely, for every $n \in \mathbb{N}$ and any $v_1, ..., v_n \in B$

$$\sum_{j=1}^{n} \alpha_j v_j = 0 \quad \text{implies} \quad \alpha_j = 0 \text{ for every } j = 1, ..., n.$$

2. For every $u \in \mathcal{V}$ there are $n \in \mathbb{N}$, $\alpha_1, ..., \alpha_n \in \mathbb{K}$, and $v_1, ..., v_n \in B$ such that

$$u = \sum_{j=1}^{n} \alpha_j v_j.$$

Comment. Not every vector space is guaranteed to possess a Hamel basis (or another type of basis that we'll discuss later). In fact, the logical statement "Every vector space has a Hamel basis" is equivalent to the Axiom of Choice [10].

Example 15. Denote the nonnegative integers by $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$ and recall the vector space

$$\mathbb{P} = \left\{ f : \mathbb{R} \to \mathbb{R} \,\middle|\, f(x) = \sum_{k=0}^{n} a_k x^k \text{ for some } n \in \mathbb{N}_0 \text{ and } a_k \in \mathbb{R},\ k = 0, ..., n \right\}.$$

For every $x \in \mathbb{R}$ and $n \in \mathbb{N}_0$, define the sequence of functions

$$f_n(x) = x^n$$

and the set

$$B = \{ f_n(x) : n \in \mathbb{N}_0 \}.$$

Then, any finite subset $B_0 = \{ f_k(x) : k \in I \subseteq \mathbb{N}_0 \text{ with } |I| < \infty \} \subseteq B$ is linearly independent. Additionally, given any $g \in \mathbb{P}$, let $m = \deg(g)$ so that there are $\alpha_0, ..., \alpha_m \in \mathbb{R}$ with

$$g(x) = \sum_{j=0}^{m} \alpha_j x^j = \sum_{j=0}^{m} \alpha_j f_j(x).$$

Thus, B is a Hamel Basis for $\mathbb{P}$.

Example 16. Similar to $\mathbb{P}$, consider the set of real-valued sequences

$$\mathbb{R}^\omega = \left\{ \{a_n\}_{n=1}^{\infty} : a_n \in \mathbb{R} \text{ for every } n \in \mathbb{N} \right\}.$$

Notice that $\mathbb{R}^\omega$ is a vector space and $\dim(\mathbb{R}^\omega) = \infty$, as no finite collection of sequences will allow one to represent all real-valued sequences. Next, define for any $k \in \mathbb{N}$, the sequence

$$e_k = (0, 0, 0, ..., 1, 0, 0, ...) \in \mathbb{R}^\omega,$$

which contains a 1 in the kth entry and zeros in all others. Represented another way, e_k is the sequence defined for every $j \in \mathbb{N}$ by

$$(e_k)_j = \begin{cases} 1, & j = k \\ 0, & j \neq k. \end{cases}$$

Let $S = \{e_k : k \in \mathbb{N}\}$, and notice that for every $n \in \mathbb{N}$ and collection of vectors $\{v_1, ..., v_n\} \subseteq S$, if $\sum_{j=1}^{n} \alpha_j v_j = 0$ where 0 represents the zero sequence, then $\alpha_1 = \cdots = \alpha_n = 0$ due to term-by-term equivalence of sequences. Hence, any finite subset $\{v_1, ..., v_n\}$ is linearly independent. However, given $\{a_k\}_{k=1}^\infty \in \mathbb{R}^\omega$ there is not necessarily $n \in \mathbb{N}$ and $\{v_1, ..., v_n\} \subseteq S$ such that

$$\{a_k\}_{k=1}^\infty = \sum_{j=1}^{n} \alpha_j v_j$$

For instance, the sequence $\{b_k\}_{k=1}^\infty$ defined by $b_k = 1$ for all $k \in \mathbb{N}$ cannot be represented as a finite linear combination of elements from S. Therefore, S is NOT a Hamel basis. In fact, $\mathbb{R}^\omega$ has no countable Hamel basis, and unfortunately, bases are not particularly useful if they are uncountable.

Comment. To summarize, we see that even with the natural coordinate mapping between $\mathbb{P}^n$ and $\mathbb{R}^{n+1}$ displayed by Example 12, the limiting spaces as $n \to \infty$, namely $\mathbb{P}$ and $\mathbb{R}^\omega$, have different properties; namely, $\mathbb{P}$ has a countable Hamel basis while $\mathbb{R}^\omega$ does not. This occurs because elements of $\mathbb{P}$ must possess a coordinate representation ending with a terminating sequence of 0's, while elements of $\mathbb{R}^\omega$ do not. Said another way, the coordinate mapping for any element of $\mathbb{P}$ would generate a sequence with only finitely many non-zero entries, and many elements of $\mathbb{R}^\omega$ possess infinitely many non-zero entries, e.g. $\{b_k\}_{k=1}^\infty$ described above.

For those vector spaces without a countable Hamel basis, this notion of a basis may not be very useful, and we can attempt to generalize it by using a norm structure on the vector space.

Definition 4.12. We say $B = \{v_n : n \in \mathbb{N}\} \subseteq \mathcal{V}$ is a **Schauder basis** for $\mathcal{V}$ if for every $u \in \mathcal{V}$ there exists a unique $\{\alpha_n\}_{n=1}^\infty \subseteq \mathbb{K}$ such that

$$u = \sum_{n=1}^{\infty} \alpha_n v_n. \tag{4.4}$$

Here, the equality of the infinite sum (4.4) is defined as

$$\lim_{N \to \infty} \left\| u - \sum_{n=1}^{N} \alpha_n v_n \right\| = 0 \tag{4.5}$$

where $\|\cdot\|$ is a norm (or a measure of length) defined on $\mathcal{V}$, which we will discuss in greater detail in the next section.

Comment. Because the definition of a Schauder basis involves an infinite sum rather than a finite sum, the ordering of the elements of the basis B are crucial to this definition, as a reordering of the elements of B may alter the corresponding limit of the sum in (4.4) or even cause it to diverge.

Example 17 (Fourier series)**.** The well-known Fourier series, which you may have encountered in a first class on Partial Differential Equations (MATH 455, here at Mines), serve as a good example of a Schauder basis. For any $a, b \in \mathbb{R}$ with $a < b$, define the vector space

$$L^2(a,b) = \left\{ f : (a,b) \to \mathbb{R} \;\middle|\; \int_a^b f(x)^2 dx < \infty \right\}.$$

Then, define the collection of functions

$$\mathcal{F} = \{1\} \bigcup \{\cos(nx) : n \in \mathbb{N}\} \bigcup \{\sin(nx) : n \in \mathbb{N}\},$$

which satisfies $\mathcal{F} \subseteq L^2(0, 2\pi)$.

Certainly, there are functions in $L^2(0, 2\pi)$ that cannot be represented as a finite linear combination of elements of $\mathcal{F}$. For instance $f(x) = x^2$ cannot be written in this manner. Thus, $\mathcal{F}$ is not a Hamel basis for $L^2(0, 2\pi)$. However, define the following norm (again, precisely defined later) - for every $f \in L^2(0, 2\pi)$, let

$$\|f\|_2 = \sqrt{\int_0^{2\pi} f(x)^2 dx}.$$

Then, $\mathcal{F}$, referred to as the **Fourier basis**, is a Schauder basis for $L^2(0, 2\pi)$ in the sense that for every $f \in L^2(0, 2\pi)$ there are unique a_0 and $a_n, b_n \in \mathbb{R}$ for every $n \in \mathbb{N}$ such that

$$\lim_{N\to\infty} \left\| f(x) - \left(a_0 + \sum_{n=1}^{N} [a_n \cos(nx) + b_n \sin(nx)] \right) \right\|_2 = 0$$

as the Fourier series of a given function $f \in L^2(0, 2\pi)$ converges to f in this sense. Additionally, the coefficients a_0, a_n, b_n are uniquely determined by the given function f and can be easily represented due to the orthogonality of the elements of $\mathcal{F}$, a notion we will also revisit in future sections.

We will provide additional and important examples of Schauder bases for L^2 once the notion of a norm has been well defined. In general, many sequence spaces, such as $\mathbb{R}^\omega$, and function spaces, such as $L^2(a, b)$, represent standard examples of infinite-dimensional vector spaces, while traditional vectors, polynomials, and matrices - spaces like $\mathbb{R}^p$, $\mathbb{C}^p$, $\mathbb{P}^n$, $\mathbb{R}^{p\times q}$, and $\mathbb{C}^{p\times q}$ - are all standard examples of finite-dimensional vector spaces.

4.4 Normed spaces

From the previous section, we see that the notion of a Hamel basis may not be useful for categorizing certain vector spaces, such as $L^2(0, 2\pi)$. Hence, the introduction of a Schauder basis may be necessary when possible, and this should lead us to study normed vector spaces.

Definition 4.13. Let $\mathcal{V}$ be a vector space. Then, a **norm** on $\mathcal{V}$ is a function that assigns to each $v \in \mathcal{V}$ a nonnegative real number denoted by $\|v\|$. This function must satisfy the following properties:

1. $\|0\| = 0$, and if $v \neq 0$ then $\|v\| > 0$.
2. For every $\alpha \in \mathbb{K}$, $v \in \mathcal{V}$, we have
$$\|\alpha v\| = |\alpha|\|v\|.$$
3. For every $u, v \in \mathcal{V}$, we have
$$\|u + v\| \leq \|u\| + \|v\|. \qquad \text{(Triangle Inequality)}$$

We will refer to vector spaces with a norm as **normed spaces**.

Example 18. Here are some common norms and normed vector spaces that you may have previously encountered:

1. $\mathcal{V} = \mathbb{R}^n$

   For any $v \in \mathbb{R}^n$, define the Euclidean norm by
$$\|v\|_2 = \sqrt{\sum_{j=1}^{n} |v_j|^2}.$$
   Similarly, for any $p \in [1, \infty)$, we can define the p-norm by
$$\|v\|_p = \left(\sum_{j=1}^{n} |v_j|^p \right)^{\frac{1}{p}}.$$
   Finally, we can also define this norm when $p = \infty$, so that
$$\|v\|_\infty = \max_{1 \leq j \leq n} |v_j|.$$

2. $\mathcal{V} = \mathbb{C}^n$

   For any $v \in \mathbb{C}^n$, define the complex Euclidean norm by
$$\|v\|_2 = \sqrt{\sum_{j=1}^{n} |v_j|^2}$$
   where $|v_j|$ represents the modulus of the entry $v_j \in \mathbb{C}$.

   As before, for any $p \in [1, \infty)$, we can define the p-norm by
$$\|v\|_p = \left(\sum_{j=1}^{n} |v_j| \right)^{\frac{1}{p}},$$
   and when $p = \infty$, we define the norm
$$\|v\|_\infty = \max_{1 \leq j \leq n} |v_j|.$$

3. $\mathcal{V} = L^2(a,b)$

   For any $f \in L^2(a,b)$, define the 2-norm by

$$\|f\|_2 = \sqrt{\int_a^b f(x)^2 \, dx}.$$

4. For any $p \in [1,\infty)$, we can define the space of functions

$$\mathcal{V} = L^p(a,b) = \left\{ f : (a,b) \to \mathbb{R} \;\middle|\; \int_a^b f(x)^p dx < \infty \right\},$$

   and the corresponding norm, called the p-norm, for any $f \in L^p(a,b)$ by

$$\|f\|_p = \left(\int_a^b f(x)^p \, dx \right)^{\frac{1}{p}}.$$

5. When $p = \infty$, we define

$$\mathcal{V} = L^\infty(a,b) = \left\{ f : (a,b) \to \mathbb{R} \;\middle|\; \|f\|_\infty < \infty \right\},$$

   where the norm here is

$$\|f\|_\infty = \operatorname*{ess\,sup}_{x \in [a,b]} |f(x)|.$$

Though we will use identical notation to describe the norms of vectors and functions (e.g., $\|\cdot\|_2$), we will differentiate them in the future merely by understanding what objects we are dealing with. Additionally, the subscript on a norm will generally be omitted whenever it can be inferred from context.

Clearly, a variety of different norms on $\mathbb{R}^n$ and $\mathbb{C}^n$ can be imposed. It should be noted that the norm is part of the vector space definition, so that using a different norm on the same set of vectors technically means the vector space is different. Of course, each of these gives rise to a different notion of length, and thus, a completely different geometry on the vector space.

Example 19. Fix $v = \begin{bmatrix} 3 \\ 4 \end{bmatrix}$ and let's compute some differing norms of this element of $\mathbb{R}^2$. In particular, we see that

$$\begin{aligned} \|v\|_2 &= \sqrt{3^2 + 4^2} = 5 \\ \|v\|_1 &= 3 + 4 = 7 \\ \|v\|_\infty &= \max\{3,4\} = 4. \end{aligned}$$

Geometrically, $\|v\|_2$ represents the radius of the circle on which the vector v terminates, while $\|v\|_\infty$ represents the projection of v onto its coordinate of greatest magnitude. Additionally, $\|v\|_1$ is the sum of the total distances traveled in all directions starting from the origin (i.e., 3 in the x direction and 4 in the y direction). For this reason and its graphical similarity to a path through city streets, the $\|\cdot\|_1$ norm is often referred to as the "taxicab norm".

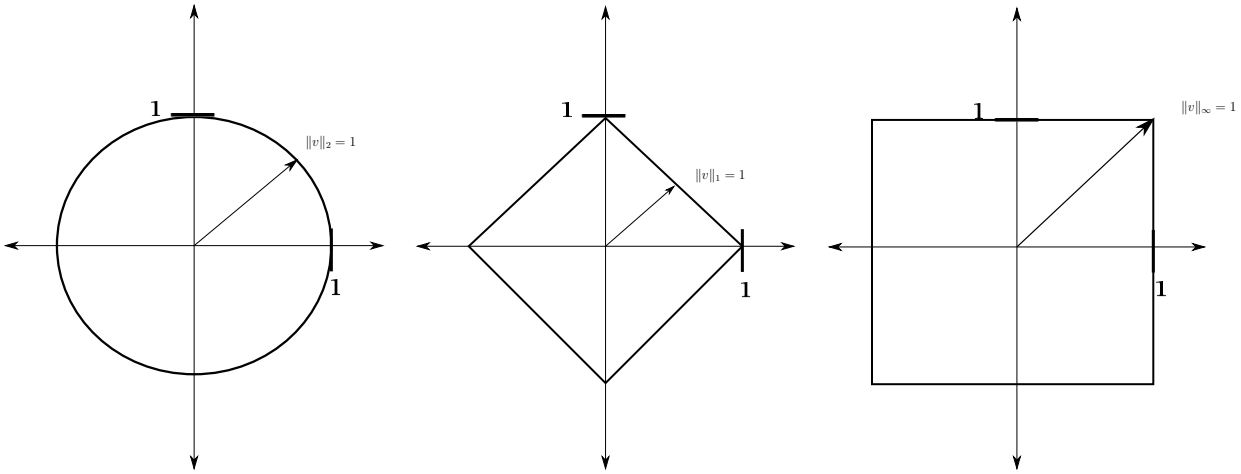


Figure 4.1: Representation of the unit ball (i.e. $\|v\| \leq 1$ with $v \in \mathbb{R}^2$) for the L^2 norm (left), L^1 norm (center), and L^∞ norm (right).

Essentially, $\|v\|_1$ treats all coordinate directions equally in determining the length of v, and as p increases, the contribution of the maximal entry of v to its length $\|v\|_p$ also increases, until we arrive at the limiting case $\|v\|_\infty$ which only considers the maximal entry of the vector v in order to define it's length (see Figure 4.1).

For normed vector spaces, the notion of an isomorphism extends in an analogous manner. In particular, two normed spaces are isomorphic if they are isomorphic as vector spaces and the norm is preserved by this mapping.

Definition 4.14. Let $\mathcal{V}$ and $\mathcal{W}$ be normed vector spaces. A mapping $T : \mathcal{V} \to \mathcal{W}$ is called an **isomorphism** if T is linear, one-to-one, onto, and

$$\|T(v)\|_{\mathcal{W}} = \|v\|_{\mathcal{V}}$$

for all $v \in \mathcal{V}$, where $\|\cdot\|_{\mathcal{V}}$ is the norm on $\mathcal{V}$ and $\|\cdot\|_{\mathcal{W}}$ is the corresponding norm on $\mathcal{W}$. In this case, we say that $\mathcal{V}$ and $\mathcal{W}$ are **isomorphic** normed spaces.

4.5 Banach spaces

Since we have now defined a notion of length on a vector space, we can discuss convergence of sequences and series in the space. This is possible because the length of the difference between a sequence (or series) and its limit is just a sequence of real numbers that tends to zero, and we know quite a bit about sequences (and series) of real numbers from Calculus.

Definition 4.15. A sequence $\{x_n\}_{n=1}^{\infty}$ (or just x_n) in a normed space $\mathcal{V}$ **converges** if there is $x \in \mathcal{V}$ such that

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

In this case, x is called the **limit** of x_n and we merely write

$$x = \lim_{n\to\infty} x_n$$

or $x_n \to x$. If x_n is not convergent, it is said to be **divergent**.

Of course, since every series can be understood merely in terms of a sequence (namely, the partial sums of the series as in (4.5)), then this definition extends to them as well. The definition above is exactly the notion of convergence used to define a Schauder basis in Definition 4.12.

It should be noted that to show the convergence of a given sequence according to this definition, one must first know its limit, and that can be quite problematic. Instead, we might try to show that the terms of the sequence ultimately get pushed closer and closer together as $n \to \infty$, and for sequences in $\mathbb{R}^n$ or $\mathbb{C}^n$ it's enough to show this property to prove convergence (see Appendix: Theorem 9.4). Unfortunately, this is not true in all normed spaces, and further analysis is needed. First, we precisely define the property that the terms of a sequence become closer together as n grows large.

Definition 4.16. A sequence $\{x_n\}_{n=1}^{\infty}$ in a normed space $\mathcal{V}$ is **Cauchy**[1] if for every $\epsilon > 0$ there is $N \in \mathbb{N}$ such that

$$\|x_m - x_n\| < \epsilon$$

for every $m, n > N$.

Thus, to summarize the above discussion, every Cauchy sequence of real (complex) vectors converges to a real (complex) vector. However, as we see in the following examples, there are Cauchy sequences in other normed spaces that do not converge to limits in the vector space.

Example 20. Consider the normed space $\mathbb{Q}$ of rational numbers with norm

$$\|x\| = |x|$$

for every $x \in \mathbb{Q}$. Next, let x_n be the sequence of rational numbers defined by

$$x_1 = 3, \quad x_2 = 3.1, \quad x_3 = 3.14, \quad x_4 = 3.141, \quad x_5 = 3.1415, \quad ...$$

so that the nth term of the sequence is the n digit truncated approximation of π. Then, x_n is Cauchy (indeed, $N \geq |\log_{10}(\epsilon)| + 1$ in the proof), but we know $x_n \to \pi \notin \mathbb{Q}$, and thus x_n does not converge to a limit in $\mathbb{Q}$.

Example 21. Consider the normed space $\mathbb{P}$ of all polynomials with norm

$$\|p\| = \max_{x \in [0,1]} |p(x)|$$

for every $p \in \mathbb{P}$. Next, consider the sequence of polynomials

$$p_n(x) = \sum_{k=0}^{n} \left(\frac{x}{2}\right)^k.$$

Then, p_n is Cauchy in this norm, as for any $m < n$ and $x \in [0, 1]$, we find

$$\|p_n(x) - p_m(x)\| = \max_{x \in [0,1]} \left| \sum_{k=m+1}^{n} \frac{x^k}{2^k} \right| \leq \sum_{k=m+1}^{n} \left(\frac{1}{2}\right)^k \leq \frac{1}{2^m} \to 0$$

[1] named after French mathematician Augustin-Louis Cauchy, who is generally considered to be one of history's most important and influential mathematicians

as $m, n \to \infty$. However, because this is a geometric series, we know

$$\lim_{n\to\infty} p_n(x) = \frac{1}{1-\frac{x}{2}} =: p_\infty(x)$$

for any $x \in [0,1]$, and this limit is not a polynomial (i.e., $p_\infty \notin \mathbb{P}$).

Since not every Cauchy sequence converges in an arbitrary normed space, we need to differentiate amongst those spaces with this property and those without.

Definition 4.17. The normed space $\mathcal{V}$ is **complete** if every Cauchy sequence in $\mathcal{V}$ converges (i.e., possesses a limit which is an element of $\mathcal{V}$). If $\mathcal{V}$ does not have this property, it is **incomplete**.

The completeness property allows normed spaces (and as we'll see in future sections, inner product spaces) to behave more like finite-dimensional vector spaces. Without this property, one can construct sequences whose limits do not belong to the vector space, which is impossible in the finite-dimensional case (see Theorem 4.12 in the next section). Thus, much of our intuition, which stems from finite-dimensional spaces, can be lost within incomplete normed spaces, and for this reason we will not focus on them.

Definition 4.18. A complete, normed vector space $\mathcal{V}$ is called a **Banach space**[2].

With this definition, we see from the previous examples that $\mathbb{Q}$ with the absolute value norm and $\mathbb{P}$ with the maximum norm are both incomplete normed spaces and thus, not Banach spaces.

Example 22. Many of the normed spaces mentioned in Example 18 are Banach spaces.

1. $\mathcal{V} = \mathbb{R}^n$ with the norm $\|v\|_p$ is a Banach space for any $p \in [1, \infty)$.
2. $\mathcal{V} = \mathbb{C}^n$ with the complex norm $\|v\|_p$ is a Banach space for any $p \in [1, \infty)$.
3. For any $p \in [1, \infty)$, the space of real sequences

$$\ell^p(\mathbb{R}) = \left\{ x = \{x_n\}_{n=1}^\infty \subset \mathbb{R} \;\middle|\; \sum_{n=1}^\infty |x_n|^p < \infty \right\}$$

   with the norm

$$\|x\| = \left(\sum_{n=1}^\infty |x_n|^p \right)^{\frac{1}{p}}$$

   is a Banach space.

4. For any $p \in [1, \infty)$, the space of functions $\mathcal{V} = L^p(a, b)$ with the p-norm is a Banach space.

[2] named after Polish mathematician Stefan Banach, who was generally considered to be one of the world's most important and influential mathematicians during the early 20th-century

5. The space of continuous functions $C[a,b]$ with the norm

$$\|f\|_\infty = \max_{x\in[a,b]} |f(x)|$$

is a Banach space

Though this last example is a Banach space, it's important to keep in mind that a different norm placed upon the same space may alter this property, as shown in the following example.

Example 23. Consider $C[a,b]$ endowed with the norm

$$\|f\|_2 = \left(\int_a^b |f(x)|^2\, dx\right)^{\frac{1}{2}}$$

instead of $\|\cdot\|_\infty$. With this norm, the space is not complete and, thus, not a Banach space. Indeed, it suffices to consider $C[0,1]$ and define the sequence

$$f_n(x) = x^n$$

for $n \in \mathbb{N}$ and $x \in [0,1]$. Then, we find

$$\begin{aligned}
\|f_n - f_m\|_2^2 &= \int_0^1 |x^n - x^m|^2\, dx = \int_0^1 \left(x^{2n} - 2x^{m+n} + x^{2m}\right)\, dx \\
&= \frac{1}{2n+1} - \frac{2}{m+n+1} + \frac{1}{2m+1} \\
&= \frac{2(n-m)^2}{(2m+1)(2n+1)(m+n)} \\
&\leq \frac{n^2+m^2-2mn}{2mn(m+n)} \\
&\leq \frac{1}{2m} + \frac{1}{2n} - \frac{1}{m+n} \to 0
\end{aligned}$$

for m, n sufficiently large. However, this sequence of continuous functions does not converge to a continuous function. Instead, taking $n \to \infty$, we see that

$$f_n(x) \to f(x) := \begin{cases} 0, & x \in [0,1) \\ 1, & x = 1 \end{cases}$$

and $f(x) \notin C[0,1]$.

Because this normed space is not complete, a natural question to ask is - can we create a Banach space out of it? The answer is, in fact, yes. In essence, we merely need to add all limits of Cauchy sequences (in the $\|\cdot\|_2$ norm) to $C[a,b]$ and this will create a new, and larger, normed vector space that will be complete, and therefore a Banach space. In this case, adding the limits of all such sequences produces exactly the Banach space $L^2[a,b]$. Similarly, performing the same procedure when using the $\|\cdot\|_p$ norm will yield $L^p[a,b]$ for any $p \in [1,\infty)$.

Since $L^2(a,b)$ arises naturally as the completion of the continuous functions in the $\|\cdot\|_2$ norm, and we discussed the Fourier basis for this space in the previous section, it makes sense to revisit this notion and describe some other bases.

Example 24 (Complex Fourier basis)**.** Similar to the Fourier series example, if we consider the functions in $L^2(0,1)$ and allow them to be complex-valued, then the collection of functions

$$\mathcal{G} = \left\{ e^{2\pi i n x} : n \in \mathbb{Z} \right\}$$

called the **complex Fourier basis**, forms a Schauder basis for the complex-valued functions in $L^2(0,1)$. These functions are often used to concisely approximate a continuous or square-integrable function in a variety of applications. Using Euler's formula, namely

$$e^{i\theta} = \cos(\theta) + i\sin(\theta),$$

with $\theta = 2\pi n x$ we can also see the connection between the basis functions of $\mathcal{G}$ and those of $\mathcal{F}$ in Example 17.

Example 25 (Haar basis)**.** Let

$$h(t) = \begin{cases} 1, & 0 < t \leq \frac{1}{2} \\ -1, & \frac{1}{2} < t \leq 1 \\ 0, & \text{else} \end{cases}$$

and for every $j, k \in \mathbb{Z}$ define

$$h_k^j(x) = 2^{j/2} h\left(2^j x - k\right),$$

which is just the function h rescaled by $2^{j/2}$ and shifted by $2^{-j}k$. Then, the collection of functions

$$\mathcal{H} = \left\{ h_k^j(x) : j, k \in \mathbb{Z} \right\},$$

called the **Haar basis** is a Schauder basis for $L^2(0,1)$. Notice that, unlike the Fourier basis, the Haar basis has **compact support** (i.e., each of the basis functions is zero outside of a closed and bounded set). Even so, any continuous function can be represented as the **uniform** limit of a linear combination of such basis functions, which is not a property enjoyed by Fourier Series. These, and many other properties, make the Haar basis particularly useful in wavelet analysis and image compression.

Comment. As we will see later, these collections of functions ($\mathcal{F}, \mathcal{G}$, and $\mathcal{H}$) each form an orthonormal basis for $L^2(0,1)$ and can be extended to $L^2(a,b)$ by scaling and translation.

Comment. For those with an understanding of stochastic processes, any of these Schauder bases for L^2 can be used to represent a (zero-mean) square-integrable stochastic process, as well. This result is referred to as the **Karhunen-Loéve theorem**, or sometimes the **KL expansion** of a random process.

Finally, we conclude this section by briefly discussing the inherent relationships between norms and related ideas concerning distance and angle.

Comment. A few, more general, comments are needed prior to ending this section:

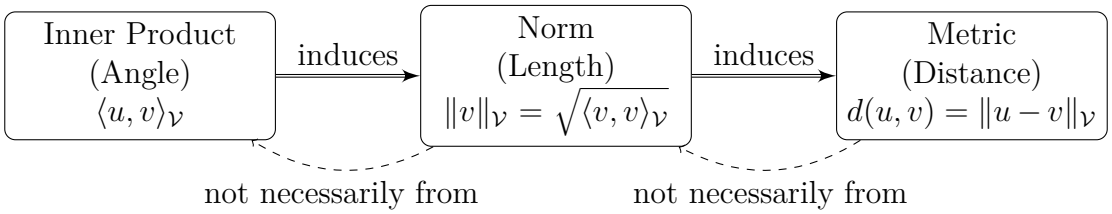


Figure 4.2: A notion of angle between vectors, called an inner product, naturally induces a notion of length of vectors, called a norm, which in turn naturally induces a notion of distance between vectors, called a metric. However, norms need not arise from inner products (e.g., $\|v\|_p$) and metrics need not arise from norms (e.g., Wasserstein metric).

1. Any vector space $\mathcal{V}$ endowed with a norm $\|\cdot\|_{\mathcal{V}}$ naturally induces a notion of distance, called a **metric** defined by

$$d(u, v) = \|u - v\|_{\mathcal{V}}.$$

   Metric spaces (i.e., vector spaces endowed with a metric) need not arise from a norm. Famous examples of such metrics include the Wasserstein metric [29] and any metric that is translation invariant [10]. In general, metric spaces are crucial to the mathematical study of Differential Geometry and its fundamental applications to General Relativity.

2. Some norms are themselves naturally induced by a notion of angle on a vector space, called an **inner product**, which we will introduce in the next section. As with metrics, not every norm arises from an inner product. In particular, on $\mathbb{R}^n$ the norm $\|\cdot\|_2$ is induced by an inner product, but $\|\cdot\|_p$ is not for $p \neq 2$.

3. Not every vector space has the property that a norm can be defined on it (e.g., non-metrizable spaces). This topic directs us toward a discussion concerning Topology, which is outside the scope of the course, but it should be mentioned that once open sets (i.e., a topology) are defined on a vector space, they need not arise from a norm (e.g., Frechet spaces).

4.6 Finite-dimensional normed spaces

In this section, we will discuss two particularly interesting and useful facts about finite-dimensional vector spaces, namely that they are all necessarily Banach spaces and that every norm placed on such a space is equivalent. First, we need a technical lemma.

Lemma 4.11. [10] Let $\mathcal{V}$ be a normed space with $\{v_1, ..., v_n\}$ a LI subset. Then, there is $C > 0$ such that for every $\alpha_1, ..., \alpha_n$

$$|\alpha_1| + \cdots + |\alpha_n| \leq C\|\alpha_1 v_1 + ...\alpha_n v_n\|.$$

Proof. Because the set $\{v_1, ..., v_n\}$ is LI, it suffices to prove this result for any $\alpha_1, ..., \alpha_n$ with $\sum_{k=1}^n |\alpha_k| = 1$. Indeed, if this sum is given to be zero, then the

linear independence of the vectors guarantees that the above inequality holds. While if this sum is not zero, then we can normalize the scalars by it, resulting in the condition $\sum_{k=1}^{n} |\alpha_k| = 1$. Hence, it suffices to take the left side of the inequality above to be 1.

Now, we proceed by contradiction. Suppose that the theorem is false, so that for any $C > 0$ there is $\beta_1, ..., \beta_n$ with $\sum_{k=1}^{n} |\beta_k| = 1$ such that

$$\frac{1}{C} \geq \|\beta_1 v_1 + ...\beta_n v_n\|.$$

Since $C > 0$ is arbitrary, we can essentially make the right side as small as desired. Thus, there is a sequence $\{u_m\}_{m=1}^{\infty}$ in $\mathcal{V}$ with

$$u_m = \alpha_1^{(m)} v_1 + \cdots + \alpha_n^{(m)} v_n$$

and $\sum_{k=1}^{n} |\alpha_k^{(m)}| = 1$ such that $\|u_m\| \to 0$ as $m \to \infty$. Because $\sum_{k=1}^{n} |\alpha_k^{(m)}| = 1$, we find $|\alpha_k^{(m)}| \leq 1$, which means that for every fixed $k = 1, ..., n$ the infinite sequence $\{\alpha_k^{(m)}\}_{m=1}^{\infty}$ is bounded. Hence, by the Bolzano-Weierstrauss Theorem (Appendix: Theorem 9.1), $\{\alpha_1^{(m)}\}_{m=1}^{\infty}$ has a convergent subsequence (say $\gamma_1^{(m)}$). Let α_1 denote the limit of this subsequence and define $\{u_{1,m}\}_{m=1}^{\infty}$ to be the subsequence of $\{u_m\}_{m=1}^{\infty}$ that corresponds to $\gamma_1^{(m)}$. By the same argument, $\{u_{1,m}\}_{m=1}^{\infty}$ has a subsequence $\{u_{2,m}\}_{m=1}^{\infty}$ for which $\{\alpha_2^{(m)}\}_{m=1}^{\infty}$ converges. Thus, we let α_2 denote the limit of this subsequence (say $\gamma_2^{(m)}$), and continue in this fashion n times to obtain a subsequence $\{u_{n,m}\}_{m=1}^{\infty}$ with terms

$$u_{n,m} = \gamma_1^{(m)} v_1 + \cdots + \gamma_n^{(m)} v_n$$

for some scalars $\gamma_k^{(m)}$ for $k = 1, ..., n$ satisfying $\sum_{k=1}^{n} |\gamma_k^{(m)}| = 1$ and $\lim_{m\to\infty} \gamma_k^{(m)} = \alpha_k$ for every $k = 1, ..., n$. Thus, we take the limit of $u_{n,m}$ as $m \to \infty$ to find

$$\lim_{m\to\infty} u_{n,m} = \lim_{m\to\infty} \left(\gamma_1^{(m)} v_1 + \cdots + \gamma_n^{(m)} v_n\right) = \sum_{k=1}^{n} \alpha_k v_k =: u.$$

Notice that the condition $\sum_{k=1}^{n} |\alpha_k| = 1$ is preserved by the limit, which means that not all of these scalars can be zero. Hence, by the LI property of $\{v_1, ..., v_n\}$, we see that $u \neq 0$. However, we also see that $\|u_{m,n}\| \to \|u\|$ as $m \to \infty$. Because $\|u_n\| \to 0$ as $n \to \infty$ by assumption and $u_{m,n}$ is merely a subsequence of u_n, we find $\|u\| = 0$, which implies $u = 0$. Hence, we arrive at a contradiction, and the proof is complete. □

With this, we may prove the completeness result.

Theorem 4.12. Let $\mathcal{V}$ be a normed space with $\dim(\mathcal{V}) < \infty$. Then, $\mathcal{V}$ is complete, and thus a Banach space.

Proof. Let $\{u_m\}_{m=1}^{\infty}$ be a Cauchy sequence in $\mathcal{V}$, define $n = \dim(\mathcal{V})$, and choose a basis of $\mathcal{V}$ denoted by $B = \{v_1, ..., v_n\}$. Then, by Theorem 4.5 each term of the sequence u_m has a unique representation with respect to B so that

$$u_m = \alpha_1^{(m)} v_1 + \cdots + \alpha_n^{(m)} v_n$$

for some $\alpha_k^{(m)} \in \mathbb{K}$ with $k = 1, ..., n$. Then, since u_m is Cauchy, for every $\epsilon > 0$ there is $N > 0$ such that

$$\|u_m - u_\ell\| < \epsilon$$

for $\ell, m > N$. By Lemma 4.11, there is $C > 0$ such that

$$|\alpha_1^{(m)} - \alpha_1^{(\ell)}| + \cdots + |\alpha_n^{(m)} - \alpha_n^{(\ell)}| \leq C \left\| \left(\alpha_1^{(m)} - \alpha_1^{(\ell)}\right) v_1 + \cdots + \left(\alpha_n^{(m)} - \alpha_n^{(\ell)}\right) v_n \right\|.$$

Thus, we find for $\ell, m > N$

$$\begin{aligned} \sum_{k=1}^{n} |\alpha_k^{(m)} - \alpha_k^{(\ell)}| &\leq C \left\| \sum_{k=1}^{n} \left(\alpha_k^{(m)} - \alpha_k^{(\ell)}\right) v_k \right\| \\ &= C \left\| \sum_{k=1}^{n} \alpha_k^{(m)} v_k - \sum_{k=1}^{n} \alpha_k^{(\ell)} v_k \right\| \\ &= C\|u_m - u_\ell\| \\ &< C\epsilon. \end{aligned}$$

Since each term of this sum is nonnegative, we can conclude

$$|\alpha_k^{(m)} - \alpha_k^{(\ell)}| \leq C\epsilon$$

for every $\ell, m > N$. Of course, this implies for every fixed $k = 1, ..., n$ the sequence $\{\alpha_k^{(m)}\}_{m=1}^{\infty}$ is a Cauchy sequence of scalars, and thus convergent. Denote the corresponding limits of these sequences by α_k for $k = 1, ..., n$ and define

$$u = \alpha_1 v_1 + \cdots + \alpha_n v_n.$$

Of course, $u \in \mathcal{V}$ by the closure properties of $\mathcal{V}$, and by the Triangle Inequality we have

$$\|u_m - u\| = \left\| \sum_{k=1}^{n} \left(\alpha_k^{(m)} - \alpha_k\right) v_k \right\| \leq \sum_{k=1}^{n} |\alpha_k^{(m)} - \alpha_k| \cdot \|v_k\| \leq M \sum_{k=1}^{n} |\alpha_k^{(m)} - \alpha_k|$$

where $M = \max_{k=1,..,n} \|v_k\|$. Because $\alpha_k^{(m)} \to \alpha_k$ as $m \to \infty$ for every $k = 1, ..., n$, we see that $\|u_m - u\| \to 0$ as $m \to \infty$. Therefore, u_m is convergent with limit $u \in \mathcal{V}$, and since u_m was an arbitrary Cauchy sequence, the completeness of $\mathcal{V}$ follows. □

Prior to establishing the second important property, we will need one more preliminary result. Recall that for any $u, v \in \mathbb{R}^n$, we use the notation $u^T v = u \cdot v$ to denote the dot product of the vectors u and v.

Theorem 4.13 (Cauchy-Schwarz Inequality for $\|\cdot\|_2$)**.** For any $u, v \in \mathbb{R}^n$, we have

$$|u^T v| \leq \|u\|_2 \|v\|_2.$$

Though this crucial result requires only an understanding of the dot product of two vectors, we will postpone its proof and instead prove it in a more general setting involving any inner product (which is a term that will be precisely defined later). Let's first demonstrate how Cauchy-Schwarz might be useful.

Example 26. We wish to verify that $\|\cdot\|_2$ is, in fact, a norm on $\mathbb{R}^n$. First, we see that $0 \in \mathbb{R}^n$ satisfies

$$\|0\|_2 = \sqrt{\sum_{j=1}^{n} 0^2} = 0.$$

Additionally, if $v \in \mathbb{R}^n$ is nonzero, then there exists $m \in \mathbb{N}$ with $1 \leq m \leq n$ such that $v_m \neq 0$. Thus,

$$\|v\|_2 = \sqrt{\sum_{j=1}^{n} |v_j|^2} \geq \sqrt{|v_m|^2} = |v_m| > 0.$$

Next, we let $\alpha \in \mathbb{R}$ and $v \in \mathbb{R}^n$ be given, and note that

$$\|\alpha v\|_2 = \sqrt{\sum_{j=1}^{n} |\alpha v_j|^2} = |\alpha| \sqrt{\sum_{j=1}^{n} |v_j|^2} = |\alpha| \|v\|_2.$$

Finally, we must show the Triangle Inequality, namely

$$\|u + v\|_2 \leq \|u\|_2 + \|v\|_2$$

for every $u, v \in \mathbb{R}^n$. Here, the Cauchy-Schwarz inequality will make life easy. Indeed, we use this theorem to find

$$\begin{aligned} \|u+v\|_2^2 &= (u+v) \cdot (u+v) \\ &= u \cdot u + v \cdot u + u \cdot v + v \cdot v \\ &= \|u\|_2^2 + 2(u \cdot v) + \|v\|_2^2 \\ &\leq \|u\|_2^2 + 2|u^T v| + \|v\|_2^2 \\ &\leq \|u\|_2^2 + 2\|u\|_2 \|v\|_2 + \|v\|_2^2 \\ &= \left(\|u\|_2 + \|v\|_2\right)^2. \end{aligned}$$

Finally, taking the square root of both sides yields the Triangle Inequality, and combining these results proves that $\|\cdot\|_2$ is a norm.

Now, we turn to the equivalence of norms on finite-dimensional spaces.

Definition 4.19. We say that two norms $\|\cdot\|_A$ and $\|\cdot\|_B$ defined on the same vector space $\mathcal{V}$ are **equivalent** if there are $C_1, C_2 > 0$ such that

$$C_1 \|v\|_A \leq \|v\|_B \leq C_2 \|v\|_A$$

for any $v \in \mathcal{V}$.

Comment. It is identical to define equivalence of two norms by the following: there exists $C > 1$ such that

$$\frac{1}{C} \|v\|_A \leq \|v\|_B \leq C \|v\|_A$$

for any $v \in \mathcal{V}$.

The idea here is not that the values of different norms are equal (as evidenced by Example 19), but instead that each norm can be controlled, i.e. bounded both above and below, by the other.

Theorem 4.14. If $\mathcal{V}$ is a normed vector space with $\dim(\mathcal{V}) < \infty$, then all norms on $\mathcal{V}$ are equivalent.

Proof. Since $\dim(\mathcal{V}) < \infty$, define $n \in \mathbb{N}$ to be the dimension of $\mathcal{V}$. Let $B = \{v_1, ..., v_n\}$ be a basis for $\mathcal{V}$ so by Lemma 4.5, for any $u \in \mathcal{V}$ there are unique $\alpha_1, ..., \alpha_n \in \mathbb{K}$ such that

$$u = \sum_{k=1}^{n} \alpha_k v_k.$$

Next, define a function $\phi : \mathcal{V} \to \mathbb{R}$ by

$$\phi(u) = \sqrt{\sum_{k=1}^{n} |\alpha_k|^2}$$

so that ϕ represents the 2-norm of u with respect to the coordinate mapping generated by the basis B. Of course, we can prove that ϕ is a norm. Since B is a basis for $\mathcal{V}$, the set $\{v_1, ..., v_n\}$ is linearly independent and thus

$$u = \sum_{k=1}^{n} \alpha_k v_k = 0 \qquad \text{iff} \qquad \alpha_k = 0 \text{ for all } k = 1, ..., n \qquad \text{iff} \qquad \phi(u) = 0.$$

Additionally, the scalar multiplication property and Triangle Inequality follow as they did for the 2-norm in Example 26. Hence, ϕ is a norm.

Now, let $\|\cdot\|$ be an arbitrary norm on $\mathcal{V}$. By the Triangle Inequality, we have for any $u \in \mathcal{V}$

$$\|u\| = \left\| \sum_{k=1}^{n} \alpha_k v_k \right\| \leq \sum_{k=1}^{n} |\alpha_k| \cdot \|v_k\|. \tag{4.6}$$

Considering the vectors of nonnegative real numbers

$$\alpha = \begin{bmatrix} |\alpha_1| \\ |\alpha_2| \\ \vdots \\ |\alpha_n| \end{bmatrix} \qquad \text{and} \qquad v = \begin{bmatrix} \|v_1\| \\ \|v_2\| \\ \vdots \\ \|v_n\| \end{bmatrix}$$

we have by Cauchy-Schwarz,

$$|\alpha^T v| \leq \|\alpha\|_2 \|v\|_2$$

or written another way

$$\sum_{k=1}^{n} |\alpha_k| \|v_k\| \leq \sqrt{\sum_{k=1}^{n} |\alpha_k|^2} \sqrt{\sum_{k=1}^{n} \|v_k\|^2}.$$

Putting this inequality together with (4.6), we find for any $u \in \mathcal{V}$

$$\|u\| \leq \sqrt{\sum_{k=1}^{n} |\alpha_k|^2} \sqrt{\sum_{k=1}^{n} \|v_k\|^2} = C_2 \phi(u) \tag{4.7}$$

where

$$C_2 = \sqrt{\sum_{k=1}^{n} \|v_k\|^2}$$

is independent of the choice of u, and is thus constant.

Now, define the set

$$S = \left\{ (\beta_1, ..., \beta_n) \in \mathbb{K}^n : \sqrt{\sum_{k=1}^{n} |\beta_k|^2} = 1 \right\}$$

and the function $f : S \to \mathbb{R}$ by

$$f(\beta) = f(\beta_1, ..., \beta_n) = \left\| \sum_{k=1}^{n} \beta_k v_k \right\|.$$

Since S is a compact (i.e., closed and bounded - see Appendix: Theorem 9.2) set and f is a continuous function, it follows from the Extreme Value Theorem (see Appendix: Theorem 9.3 or [21]) that f must attain a minimum on S. Thus, there are $\gamma_1, ..., \gamma_n \in \mathbb{K}$ such that

$$f(\gamma_1, ..., \gamma_n) = \min_{\beta_1, ..., \beta_n \in S} f(\beta_1, ..., \beta_n) = C_1$$

with

$$\sqrt{\sum_{k=1}^{n} |\gamma_k|^2} = 1.$$

Then, for any $u \in \mathcal{V}$ we have

$$\begin{aligned}
\|u\| &= \left\| \sum_{k=1}^{n} \alpha_k v_k \right\| \\
&= \frac{\sqrt{\sum_{k=1}^{n} |\alpha_k|^2}}{\sqrt{\sum_{k=1}^{n} |\alpha_k|^2}} \cdot \left\| \sum_{k=1}^{n} \alpha_k v_k \right\| \\
&= \sqrt{\sum_{k=1}^{n} |\alpha_k|^2} \cdot \left\| \sum_{k=1}^{n} \frac{\alpha_k}{\sqrt{\sum_{k=1}^{n} |\alpha_k|^2}} v_k \right\| \\
&\geq \sqrt{\sum_{k=1}^{n} |\alpha_k|^2} \left\| \sum_{k=1}^{n} \gamma_k v_k \right\| \\
&= C_1 \sqrt{\sum_{k=1}^{n} |\alpha_k|^2} \\
&= C_1 \phi(u).
\end{aligned}$$

Finally, combining this with the previously-established inequality (4.7), we have

$$C_1 \phi(u) \leq \|u\| \leq C_2 \phi(u)$$

for every $u \in \mathcal{V}$, and thus the norms are equivalent. Since $\|\cdot\|$ was arbitrary, we see that all norms defined on $\mathcal{V}$ are equivalent, and the proof is complete. □

4.7 Inner product spaces & Hilbert spaces

Now that we better understand the norm structure on a vector space, we can generalize this idea to an inner product, which defines the notion of angles between vectors.

Definition 4.20. Let $\mathcal{V}$ be a vector space. An **inner product** (dot product or scalar product) on $\mathcal{V}$ is a function that assigns to each pair $u, v \in \mathcal{V}$ a scalar in $\mathbb{K}$, denoted by $\langle u, v\rangle$, satisfying the following properties

1. $\langle 0, 0\rangle = 0$ and if $v \neq 0$ then $\langle v, v\rangle > 0$
2. Conjugate Symmetry:
   For every $u, v \in \mathcal{V}$,
   $$\langle u, v\rangle = \overline{\langle v, u\rangle}$$
   and if $\mathbb{K} = \mathbb{R}$, then
   $$\langle u, v\rangle = \langle v, u\rangle.$$
3. Conjugate Linearity in the first argument:
   For every $\alpha, \beta \in \mathbb{K}$ and $u, v, w \in \mathcal{V}$,
   $$\langle \alpha u + \beta v, w\rangle = \overline{\alpha}\langle u, w\rangle + \overline{\beta}\langle v, w\rangle$$
   and if $\mathbb{K} = \mathbb{R}$, then
   $$\langle \alpha u + \beta v, w\rangle = \alpha\langle u, w\rangle + \beta\langle v, w\rangle.$$

A vector space endowed with such an inner product is referred to as an **inner product space**.

Comment. As previously mentioned, every inner product induces a norm on $\mathcal{V}$. In particular, this induced norm is defined by

$$\|v\|_{\mathcal{V}} = \sqrt{\langle v, v\rangle}$$

for every $v \in \mathcal{V}$. The first two norm properties can be verified directly from the properties of the inner product, while the third requires a general version of the Cauchy-Schwarz inequality that we will state and prove shortly.

Comment. Notice that for $\mathbb{K} = \mathbb{C}$, the conjugation within the linearity property only occurs in the first entry of the inner product, so that

$$\langle w, \alpha u + \beta v\rangle = \alpha\langle w, u\rangle + \beta\langle w, v\rangle.$$

Indeed, using the properties of the inner product, we see

$$\begin{aligned}\langle w, \alpha u + \beta v\rangle &= \overline{\langle \alpha u + \beta v, w\rangle} \\ &= \overline{\overline{\alpha}\langle u, w\rangle + \overline{\beta}\langle v, w\rangle} \\ &= \alpha\overline{\langle u, w\rangle} + \beta\overline{\langle v, w\rangle} \\ &= \alpha\langle w, u\rangle + \beta\langle w, v\rangle.\end{aligned}$$

Of course, this equality holds in the case $\mathbb{K} = \mathbb{R}$ as well.

Comment. In this same case (i.e. $\mathbb{K} = \mathbb{C}$), an equivalent definition of inner product can be generated by replacing the conjugate linearity condition with standard linearity, namely, for every $\alpha, \beta \in \mathbb{K}$ and $u, v, w \in \mathcal{V}$,

$$\langle \alpha u + \beta v, w \rangle = \alpha \langle u, w \rangle + \beta \langle v, w \rangle.$$

If this is done, it follows from the previous comment that the inner product is then *conjugate* linear in the second (rather than the first) argument. Of course, this is merely a matter of convention and while some sources use the latter formulation, we will stick to Definition 4.20 so as to utilize the familiar Hermitian transpose notation as in the second example below.

Example 27. Here are some prominent examples of inner products on vector spaces, a few of which we've already touched on.

1. $\mathcal{V} = \mathbb{R}^p$

   For any $u, v \in \mathbb{R}^p$, the dot product is an inner product, which is often called the standard inner product on $\mathbb{R}^p$, defined by

   $$\langle u, v \rangle = \sum_{j=1}^{p} u_j v_j = u \cdot v = u^T v.$$

   The norm induced by this inner product is the **Euclidean norm**, which is defined for any $u \in \mathbb{R}^p$ by

   $$\|u\|_2 = \sqrt{\langle u, u \rangle} = \sqrt{\sum_{j=1}^{p} |u_j|^2}.$$

2. $\mathcal{V} = \mathbb{C}^p$

   For any $u, v \in \mathbb{C}^p$, the conjugate dot product is an inner product, which is often called the standard inner product on $\mathbb{C}^p$, defined by

   $$\langle u, v \rangle = \sum_{j=1}^{p} \overline{u}_j v_j = u^H v.$$

   The norm induced by this inner product is defined for any $u \in \mathbb{C}^p$ by

   $$\|u\|_2 = \sqrt{\langle u, u \rangle} = \sqrt{\sum_{j=1}^{p} |u_j|^2}.$$

3. $\mathcal{V} = L^2(a, b)$

   For any $f, g \in L^2(a, b)$, we can define an inner product by

   $$\langle f, g \rangle = \int_a^b f(x) g(x) \, dx.$$

   The norm induced by this inner product is defined for any $f \in L^2(a, b)$ by

   $$\|f\|_2 = \sqrt{\langle f, f \rangle} = \sqrt{\int_a^b |f(x)|^2 \, dx}.$$

4. $\mathcal{V} = \mathbb{R}^{p\times q}$

   For any $A, B \in \mathbb{R}^{p\times q}$, we may define an inner product, called the **Frobenius inner product**, by

   $$\langle A, B\rangle = \operatorname{tr}\left(A^T B\right).$$

   The norm induced by this inner product is called the **Frobenius norm** and is defined for any $A \in \mathbb{R}^{p\times q}$ by

   $$\|A\|_F = \sqrt{\langle A, A\rangle} = \sqrt{\operatorname{tr}\left(A^T A\right)} = \sqrt{\sum_{j=1}^{q} \|a_j\|_2^2} = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} |a_{ij}|^2}$$

   where a_j represents the jth column of A and a_{ij} is the (i,j)th entry of A.

5. The space of real sequences

   $$\ell^2(\mathbb{R}) = \left\{ x = \{x_n\}_{n=1}^{\infty} \subset \mathbb{R} \;\middle|\; \sum_{n=1}^{\infty} |x_n|^2 < \infty \right\}$$

   with the inner product

   $$\langle x, y\rangle = \sum_{n=1}^{\infty} x_n y_n$$

   is an inner product space.

6. None of the other normed spaces mentioned in Example 18 are inner product spaces.

For the remainder of the section, we will tacitly assume that $\mathcal{V}$ is a vector space with a given inner product denoted by $\langle\cdot,\cdot\rangle$ and $\|\cdot\|$ is the associated norm induced by this inner product.

For inner product spaces, the notion of an isomorphism extends in an analogous manner, as well. In particular, two inner product spaces are isomorphic if they are isomorphic as vector spaces and the inner product is preserved by this mapping.

Definition 4.21. Let $\mathcal{V}$ and $\mathcal{W}$ be inner product spaces. A mapping $T : \mathcal{V} \to \mathcal{W}$ is called an **isomorphism** if T is linear, one-to-one, onto, and

$$\langle T(u), T(v)\rangle_{\mathcal{W}} = \langle u, v\rangle_{\mathcal{V}}$$

for all $u, v \in \mathcal{V}$ where $\langle\cdot,\cdot\rangle_{\mathcal{V}}$ is the inner product on $\mathcal{V}$ and $\langle\cdot,\cdot\rangle_{\mathcal{W}}$ is the corresponding inner product on $\mathcal{W}$. In this case, we say that $\mathcal{V}$ and $\mathcal{W}$ are **isomorphic** inner product spaces.

Next, we prove some basic properties of inner product spaces.

Lemma 4.15. For any $v \in \mathcal{V}$,

$$\langle 0, v\rangle = \langle v, 0\rangle = 0.$$

Proof. Let $v \in \mathcal{V}$ be given. Then, by the properties of the inner product we have

$$\langle 0, v\rangle = \langle 0 \cdot v, v\rangle = 0\langle v, v\rangle = 0.$$

Finally, the conjugate symmetry of the inner product further implies

$$\langle v, 0\rangle = \overline{\langle 0, v\rangle} = \overline{0} = 0.$$

□

Theorem 4.16 (Cauchy-Schwarz Inequality)**.** For every $u, v \in \mathcal{V}$

$$|\langle u, v\rangle|^2 \leq \langle u, u\rangle \cdot \langle v, v\rangle$$

or stated another way,

$$|\langle u, v\rangle| \leq \|u\| \cdot \|v\|.$$

Proof of Cauchy-Schwarz. Let $u, v \in \mathcal{V}$ be given. If $v = 0$, then both sides of the inequality are necessarily zero by Lemma 4.15, and therefore the result holds. Now, if $v \neq 0$ we define

$$w = u - \frac{\langle v, u\rangle}{\langle v, v\rangle} v.$$

Intuitively, w represents the projection of u onto the hyperplane orthogonal to v. Then, we see that

$$\langle v, w\rangle = \langle v, u\rangle - \frac{\langle v, u\rangle}{\langle v, v\rangle}\langle v, v\rangle = 0.$$

Additionally, from the definition of w we can write

$$u = w + \frac{\langle v, u\rangle}{\langle v, v\rangle} v.$$

Using the fact that $\langle v, w\rangle = 0$ (which implies $\langle w, v\rangle = 0$) and $\langle w, w\rangle \geq 0$ we have

$$\begin{aligned}
\langle u, u\rangle &= \left\langle w + \frac{\langle v, u\rangle}{\langle v, v\rangle} v, w + \frac{\langle v, u\rangle}{\langle v, v\rangle} v \right\rangle \\
&= \langle w, w\rangle + \frac{\langle v, u\rangle}{\langle v, v\rangle}\langle w, v\rangle + \frac{\overline{\langle v, u\rangle}}{\langle v, v\rangle}\langle v, w\rangle + \frac{|\langle v, u\rangle|^2}{|\langle v, v\rangle|^2}\langle v, v\rangle \\
&= \langle w, w\rangle + \frac{|\langle v, u\rangle|^2}{|\langle v, v\rangle|^2}\langle v, v\rangle \\
&= \langle w, w\rangle + \frac{|\langle v, u\rangle|^2}{\langle v, v\rangle} \\
&\geq \frac{|\langle v, u\rangle|^2}{\langle v, v\rangle} \\
&= \frac{|\langle u, v\rangle|^2}{\langle v, v\rangle}.
\end{aligned}$$

Finally, multiplying both sides by $\langle v, v\rangle$ yields the desired inequality. □

Another useful property in these spaces is the continuity of the inner product as shown in the following theorem.

Theorem 4.17. For every $u_n, v_n \in \mathcal{V}$ with $u_n \to u$ and $v_n \to v$ in $\mathcal{V}$, we have

$$\langle u_n, v_n \rangle \to \langle u, v \rangle.$$

Proof. Using the triangle inequality and Cauchy-Schwarz, we find

$$\begin{aligned} |\langle u_n, v_n \rangle - \langle u, v \rangle| &= |\langle u_n, v_n \rangle - \langle u_n, v \rangle + \langle u_n, v \rangle - \langle u, v \rangle| \\ &\leq |\langle u_n, v_n - v \rangle| + |\langle u_n - u, v \rangle| \\ &\leq \|u_n\| \cdot \|v_n - v\| + \|u_n - u\| \cdot \|v\| \end{aligned}$$

Now, taking $n \to \infty$, we find $\|v_n - v\| \to 0$ and $\|u_n - u\| \to 0$ by assumption. Hence, the above inequality shows

$$|\langle u_n, v_n \rangle - \langle u, v \rangle| \to 0$$

as $n \to \infty$, and thus

$$\langle u_n, v_n \rangle \to \langle u, v \rangle.$$

□

Now that we've covered the basic properties of inner product spaces, we can further discuss their completeness properties. First, we can immediately deduce that, similar to normed spaces, inner product spaces may fail to be complete.

Comment. The vector space $C[a, b]$ endowed with the inner product

$$\langle f, g \rangle = \int_a^b f(x)g(x)\, dx$$

induces the norm $\|\cdot\|_2$ and, by Example 23, must be incomplete.

Hence, we may discuss the notion of completeness in this context, as well.

Definition 4.22. A complete, inner product space $\mathcal{V}$ is called a **Hilbert space**[3].

Comment. Because inner product spaces are automatically normed, we see that all Hilbert spaces must be Banach spaces. Additionally, all of the inner product spaces mentioned in Example 27 are Hilbert spaces, as well. Finally, by Theorem 4.12, every finite dimensional inner product space is necessarily complete, and thus a Hilbert space.

To close this section, we will prove one final result that highlights the need for the completeness property and will prove very useful in subsequent sections. First, we require a definition.

Definition 4.23. A subset M of a normed space $\mathcal{V}$ is **closed** if for every sequence $\{v_n\}_{n=1}^{\infty} \subset M$ with $v_n \to v$ for some $v \in \mathcal{V}$, we have $v \in M$. Said another way, M is closed if it contains the limits of all of its convergent sequences.

Theorem 4.18. Let $\mathcal{V}$ be a Hilbert space and $M \subseteq \mathcal{V}$ be a closed subspace. Then, M is complete.

[3]named after German mathematician David Hilbert, who is generally considered to be one of the most important and influential mathematicians of all time

Proof. To prove that M is complete, we let $v_n \in M$ for $n \in \mathbb{N}$ be a Cauchy sequence. Then, because $M \subseteq \mathcal{V}$, we see that $v_n \in \mathcal{V}$ for every $n \in \mathbb{N}$. Since $\mathcal{V}$ is complete, this Cauchy sequence of elements from $\mathcal{V}$ must converge to a limit $v \in \mathcal{V}$. Finally, because M is closed, it must contain all of its limit points, or said another way, the limit of every convergent sequence in M must also be in M. Hence, $v \in M$, and since v_n was an arbitrary Cauchy sequence that has been shown to converge to a limit in M, the subspace M is complete. □

4.8 Orthogonality and its Consequences

Next, we use the inner product to precisely define the notion of orthogonal elements of a vector space. Throughout, we assume $\mathcal{V}$ is a Hilbert space.

Definition 4.24. Let I be an index set that could be, for instance $I = \{1, ..., n\}$ for some $n \in \mathbb{N}$ or $I = \mathbb{N}$. We say

1. The elements $u, v \in \mathcal{V}$ are **orthogonal** (denoted $u \perp v$) if
$$\langle u, v \rangle = 0.$$
2. The set $S = \{v_j\}_{j \in I} \subseteq \mathcal{V}$ is **orthogonal** if $\langle v_j, v_k \rangle = 0$ for every $j, k \in I$ with $j \neq k$.
3. The set $S = \{v_k\}_{k \in I} \subseteq \mathcal{V}$ is **orthonormal** if S is orthogonal and $\|v_j\| = 1$ for every $j \in I$.

Example 28. Here are a few examples of orthogonality in Hilbert spaces.

1. Let
$$u = \begin{bmatrix} 3 \\ 0 \\ -1 \end{bmatrix} \qquad \text{and} \qquad v = \begin{bmatrix} -1 \\ 5 \\ -3 \end{bmatrix}.$$
Then, these vectors satisfy
$$\langle u, v \rangle = u^T v = -3 + 3 = 0,$$
and thus are orthogonal with respect to the standard inner product on $\mathbb{R}^3$. As neither is a unit vector, the set $S = \{u, v\}$ is orthogonal but not orthonormal.

2. Let
$$A = \begin{bmatrix} 1 & -2 \\ 1 & 4 \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} 1 & -2 \\ -1 & -1 \end{bmatrix}$$
and note that
$$A^T B = \begin{bmatrix} 1 & 1 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} 1 & -2 \\ -1 & -1 \end{bmatrix} = \begin{bmatrix} 0 & -3 \\ -6 & 0 \end{bmatrix}.$$
Therefore, $\text{tr}(A^T B) = 0$, and we see that A and B are orthogonal with respect to the Frobenius inner product on $\mathbb{R}^{2\times 2}$. In particular, we note that each of the columns of A are orthogonal to their respective counterparts in B. Said another way, if we denote $A = [a_1 \ a_2]$ and $B = [b_1 \ b_2]$, then
$$a_1 \cdot b_1 = a_2 \cdot b_2 = 0.$$

3. Let $m, n \in \mathbb{N}$ be given and define

$$f(x) = \sin(nx) \qquad \text{and} \qquad g(x) = \cos(mx).$$

Then, $f, g \in L^2(0, 2\pi)$ are orthogonal with respect to the standard inner product on this space; that is

$$\int_0^{2\pi} f(x)g(x)\, dx = 0,$$

and this remains true for any choice of $m, n \in \mathbb{N}$.

As we have seen with traditional vectors, orthogonal subsets of a vector space that do not contain the zero vector are necessarily linearly independent.

Theorem 4.19. Let $n \in \mathbb{N}$ be given and assume $S = \{v_1, ..., v_n\} \subseteq \mathcal{V}$ is orthogonal with $0 \notin S$. Then, S is linearly independent.

Proof. The proof is assigned as a homework problem (cf. Problem 4.13). □

Definition 4.25. Due to the differences between finite- and infinite-dimensional spaces, we require two definitions for bases.

1. If $\dim(\mathcal{V}) = n \in \mathbb{N}$, then the set $B = \{v_1, ..., v_n\}$ is an **orthonormal basis** for $\mathcal{V}$ if B is orthonormal and spans $\mathcal{V}$.

2. If $\dim(\mathcal{V}) = \infty$ (and $\mathcal{V}$ is separable[4]), then the set $B = \{v_n : n \in \mathbb{N}\}$ is an **orthonormal basis** for $\mathcal{V}$ if B is orthonormal and a Schauder basis for $\mathcal{V}$.

Example 29. We've seen a few examples of orthonormal bases.

1. For $n \in \mathbb{N}$, the set $B = \{e_1, ..., e_n\}$ is an orthonormal basis for $\mathbb{R}^n$.

2. Of course, there are other orthonormal bases for these spaces. For example,

$$B = \left\{ \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \right\}$$

is an orthonormal basis for $\mathbb{R}^2$.

3. The set of functions

$$B = \left\{ \frac{1}{\sqrt{2\pi}} \right\} \cup \left\{ \frac{1}{\sqrt{\pi}} \sin(nx) : n \in \mathbb{N} \right\} \cup \left\{ \frac{1}{\sqrt{\pi}} \cos(mx) : m \in \mathbb{N} \right\}$$

is an orthonormal basis for $L^2(0, 2\pi)$.

Why are orthogonal and orthonormal bases so important?
For starters, they make a number of computations in vector spaces considerably easier than using bases without this property. In particular, orthogonal bases allow us to find the coordinates of any vector in the space without much work. Indeed, if

[4] We will not discuss separability here, but this condition is technically required for the existence of a countable Schauder basis.

we let $B = \{v_1, ..., v_n\}$ be an orthogonal basis for a Hilbert space $\mathcal{V}$, then by Lemma 4.5 for any $v \in \mathcal{V}$ there are unique $\alpha_1, ..., \alpha_n \in \mathbb{K}$ such that $v = \sum_{j=1}^{n} \alpha_j v_j$. Therefore, for any $k = 1, ..., n$, we find

$$\langle v_k, v \rangle = \left\langle v_k, \sum_{j=1}^{n} \alpha_j v_j \right\rangle = \sum_{j=1}^{n} \alpha_j \langle v_k, v_j \rangle.$$

Since the basis is orthogonal, we see that $\langle v_k, v_j \rangle = 0$ for $j \neq k$, and the relationship above reduces to

$$\langle v_k, v \rangle = \alpha_k \langle v_k, v_k \rangle.$$

Finally, since $v_k \neq 0$, we can divide both sides to find

$$\alpha_k = \frac{\langle v_k, v \rangle}{\langle v_k, v_k \rangle},$$

and this provides an explicit expression for the coordinates of any element $v \in \mathcal{V}$ with respect to a given orthogonal basis. Furthermore, if B is orthonormal, then $\langle v_k, v_k \rangle = 1$ and this yields

$$\alpha_k = \langle v_k, v \rangle.$$

Thus, the coordinates of a given vector are merely the inner products of the vector with the orthonormal basis elements. In the infinite-dimensional case, Theorem 4.17 allows us to exchange limits inside and outside of the inner product, and thus we have proved the following result.

Theorem 4.20. Let $\{v_j\}_{j \in I}$ be an orthonormal basis for $\mathcal{V}$.

1. If $I = \{1, ..., n\}$, then every $v \in \mathcal{V}$ can be represented as

$$v = \sum_{j=1}^{n} \langle v_j, v \rangle v_j.$$

2. If $I = \mathbb{N}$, then every $v \in \mathcal{V}$ can be represented as

$$v = \sum_{j=1}^{\infty} \langle v_j, v \rangle v_j.$$

Since they can be so useful, we would like to know how to construct such bases. Fortunately, a specific algorithm, known as the **Gram-Schmidt process**, has been developed to obtain an orthonormal (or just orthogonal) basis from any spanning set of an inner product space. Given a spanning set consisting of n elements, the algorithm requires n steps to construct an orthonormal basis. In general, the idea is that we may sequentially redefine elements of a spanning set in order to remove any previously-occurring directions from each vector so that the newly resulting elements are orthogonal. For instance, consider that we are given a basis for $\mathbb{R}^2$ defined by $B = \{v_1, v_2\}$ where

$$v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad v_2 = \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Then, we can define a new basis by first keeping v_1 as the initial element, and then removing any portion of v_2 that points in the v_1 direction. In this case, because v_1 merely represents the first entry of a given vector, we see that a scaling factor of $3v_1$ is included within v_2. Thus, we let $u_1 = v_1$ and then define

$$u_2 = v_2 - 3u_1 = v_2 - 3v_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Finally, the resulting set

$$\{u_1, u_2\} = \left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix} \right\}$$

is an orthogonal basis for $\mathbb{R}^2$. Generally, normalization is then required to further construct an orthonormal basis.

Theorem 4.21 (Gram-Schmidt Process)**.** Let $n \in \mathbb{N}$ be given and $S = \{v_1, ..., v_n\}$ be a spanning set for a subspace $M \subseteq \mathcal{V}$. Then, there is $k \in \mathbb{N}$ with $k \leq n$ and $w_1, ..., w_k \in \mathcal{V}$ such that $B_w = \{w_1, ..., w_k\}$ is an orthonormal basis for M.

Gram-Schmidt Algorithm. For simplicity, we will first reduce S to a basis for M, because if it is only a spanning set, then we may delete any vectors that are linear combinations of others in S (just as in Theorem 4.4) to arrive at a basis B_v consisting of $k \leq n$ elements. In particular, those vectors which can be expressed as a linear combination of others in S will result in 0 vectors under the Gram-Schmidt process, and we merely omit these from the newly-constructed orthogonal basis.

Given B_v, we begin by defining $u_1 = v_1$. Then, for every $i, j = 1, ..., k$ with $i < j$ we let

$$\alpha_{ij} = \frac{\langle u_i, v_j \rangle}{\langle u_i, u_i \rangle}.$$

Here, the u vectors will be defined sequentially by the algorithm and within the definition of a particular u_j, we may utilize the previous coefficients α_{ij} for $i < j$. Next, we let

$$\begin{aligned} u_2 &= v_2 - \alpha_{12} u_1, \\ u_3 &= v_3 - \alpha_{13} u_1 - \alpha_{23} u_2, \end{aligned}$$

and continue this process for every $j \leq k$ by letting

$$u_j = v_j - \sum_{i=1}^{j-1} \alpha_{ij} u_i.$$

Then, the set $B_u = \{u_1, ..., u_k\}$ is orthogonal by construction and must span M (which can be shown directly using induction), which means it is an orthogonal basis for M. Finally, we normalize each element so that

$$w_j = \frac{u_j}{\|u_j\|}$$

for every $j = 1, ..., k$ to obtain an orthonormal basis $B_w = \{w_1, ..., w_k\}$. □

The next example should demonstrate that any spanning set for a subspace M can be transformed into an orthogonal basis for M via Gram-Schmidt.

Example 30. Let $\mathcal{V} = \mathbb{R}^3$ and define $M = \text{span}\{v_1, v_2, v_3\}$ where

$$v_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \qquad v_2 = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}, \qquad v_3 = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix}.$$

Note that these vectors do not form a basis for M as they are linearly dependent; in particular, $v_2 = 2v_1$. We construct an orthogonal basis for M as follows. First, let $u_1 = v_1$. Then, compute

$$\alpha_{12} = \frac{\langle u_1, v_2 \rangle}{\langle u_1, u_1 \rangle} = \frac{6}{3} = 2$$

and let

$$u_2 = v_2 - \alpha_{12} u_1 = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix} - 2 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = 0.$$

Since $u_2 = 0$, we discard it and remove it from the orthogonal basis that we're constructing. Next, compute

$$\alpha_{13} = \frac{\langle u_1, v_3 \rangle}{\langle u_1, u_1 \rangle} = \frac{3}{3} = 1$$

and let

$$u_3 = v_3 - \alpha_{13} u_1 = \begin{bmatrix} 3 \\ 0 \\ 0 \end{bmatrix} - 1 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix}.$$

Thus,

$$B_u = \{u_1, u_3\} = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix} \right\}$$

is an orthogonal basis for M. To construct an orthonormal basis, we merely divide each vector by its respective length. In particular, normalizing u_1 and u_3, we let

$$w_1 = \frac{u_1}{\|u_1\|_2} = \frac{1}{\sqrt{3}} \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

and

$$w_3 = \frac{u_3}{\|u_3\|_2} = \frac{1}{\sqrt{6}} \begin{bmatrix} 2 \\ -1 \\ -1 \end{bmatrix}.$$

Then, $\{w_1, w_3\}$ is an orthonormal basis for M.

Though our examples generally come from $\mathbb{R}^n$, it should be noted that the Gram-Schmidt process is not restricted to finite-dimensional Hilbert spaces, and is often used to construct orthonormal bases for finite-dimensional *subspaces* of infinite-dimensional vector spaces. For instance, if one wishes to computationally represent a function $f \in L^2(a, b)$, they may use a specific set of functions to approximate f, but because orthonormal functions are easier to work with, they

may construct an orthonormal basis for the approximating subspace using Gram-Schmidt and then merely express the coordinates of f relative to this basis. Computationally-speaking, if one fixes an orthonormal basis, then it is much easier to store the coordinates of f (which is just an n-dimensional vector where n is the number of basis elements) relative to this basis rather than attempting to store f using a partitioned discretization of the interval $[a, b]$.

That being said, for finite-dimensional vector spaces, the Gram-Schmidt process is inherently related to a specific matrix factorization, called the **QR Factorization** of a given matrix $A \in \mathbb{R}^{p\times q}$.

Theorem 4.22 (QR Factorization)**.** For any $A \in \mathbb{R}^{p\times q} \setminus \{0\}$, there are $k \in \mathbb{N}$ with $k \leq q$, a matrix $Q \in \mathbb{R}^{p\times k}$ with orthonormal columns, and an upper triangular matrix $R \in \mathbb{R}^{k\times q}$ with $\text{rank}(R) = k$ such that $A = QR$.

Proof. Let $A \in \mathbb{R}^{p\times q} \setminus \{0\}$ be given and denote its q column vectors in $\mathbb{R}^p$ by $v_1, ..., v_q$. Then, define $M = \text{span}\{v_1, ..., v_q\}$, and use the Gram-Schmidt process to construct an orthogonal basis for M with any associated zero vectors removed, resulting in a set of $k \leq q$ vectors denoted by $\{u_1, ..., u_k\}$. Next, normalize these vectors to form an orthonormal basis for M, denoted by $B = \{w_1, ..., w_k\}$ and create the matrix $Q \in \mathbb{R}^{p\times k}$ with columns $w_1, ..., w_k$. By construction, the columns of Q are orthonormal. Finally, use the coefficients of the Gram-Schmidt process to construct $R \in \mathbb{R}^{k\times q}$ with the associated $p - k$ rows removed from R whose indices correspond to the zero column vectors removed from Q. In particular, since α_{ij} is only defined for $i < j$, the matrix R is first defined by

$$R_{ij} = \begin{cases} \|u_i\|_2\alpha_{ij}, & \text{if } i < j \\ \|u_i\|_2, & \text{if } i = j \\ 0, & \text{if } i > j \end{cases}$$

for $i = 1, ..., p$, $j = 1, ..., q$. Then, $p - k$ rows are removed from R, which correspond to any zero columns encountered during the process of building Q, to define the entries R_{ij} for $i = 1, ..., k$, $j = 1, ..., q$. Finally, it follows from this construction that R is upper triangular with $\text{rank}(R) = k$.

In practice, the decomposition takes the form

$$A = \left[\begin{array}{c|c|c} v_1 & \cdots & v_q \end{array}\right] = \underbrace{\left[\begin{array}{c|c|c} w_1 & \cdots & w_k \end{array}\right]}_{Q} \underbrace{\begin{bmatrix} \|u_1\|_2 & \|u_1\|_2\alpha_{12} & \cdots & \|u_1\|_2\alpha_{1q} \\ 0 & \|u_2\|_2 & \ddots & \vdots \\ \vdots & & & \|u_{k-1}\|_2\alpha_{(q-1)q} \\ 0 & 0 & \cdots & \|u_k\|_2 \end{bmatrix}}_{R}.$$

Note that, due to the normalization of the columns of Q, the $\|u_i\|_2$ terms for $i = 1, ..., k$ are always along the diagonal of R, but neither Q nor R need be square matrices. □

Comment. The factorization of A in the theorem is more specifically referred to as the **normalized QR Factorization** of A, and is not necessarily unique. Further, we mention that this factorization holds for complex-valued matrices, as well, with the resulting factors Q and R also possibly complex-valued.

Example 31. Continuing our previous example for Gram-Schmidt and extending it to the QR Factorization, we consider the matrix

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 0 \\ 1 & 2 & 0 \end{bmatrix}.$$

Then, we can invert the relationship between the v and u vectors from the Gram-Schmidt process. Namely, solving for the v vectors in terms of the u vectors, we find

$$\begin{aligned} u_1 &= v_1 & & & v_1 &= u_1, \\ u_2 &= v_2 - \alpha_{12} u_1 & \implies & & v_2 &= u_2 + \alpha_{12} u_1, \\ u_3 &= v_3 - \alpha_{13} u_1 - \alpha_{23} u_2 & & & v_3 &= u_3 + \alpha_{13} u_1 + \alpha_{23} u_2. \end{aligned}$$

Using this, we can decompose A as

$$\begin{aligned} A &= \left[\begin{array}{c|c|c} & & \\ v_1 & v_2 & v_3 \\ & & \end{array}\right] \\ &= \left[\begin{array}{c|c|c} & & \\ u_1 & \alpha_{12} u_1 + u_2 & \alpha_{13} u_1 + \alpha_{23} u_2 + u_3 \\ & & \end{array}\right] \\ &= \left[\begin{array}{c|c|c} & & \\ u_1 & u_2 & u_3 \\ & & \end{array}\right] \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} \\ 0 & 1 & \alpha_{23} \\ 0 & 0 & 1 \end{bmatrix} \end{aligned}$$

However, since $u_2 = 0$, we remove the second column of the matrix on the left and the second row of the matrix on the right, resulting in

$$\begin{aligned} A &= \left[\begin{array}{c|c} & \\ u_1 & u_3 \\ & \end{array}\right] \begin{bmatrix} 1 & \alpha_{12} & \alpha_{13} \\ 0 & 0 & 1 \end{bmatrix} \\ &= \left[\begin{array}{c|c} & \\ \frac{u_1}{\|u_1\|_2} & \frac{u_3}{\|u_3\|_2} \\ & \end{array}\right] \begin{bmatrix} \|u_1\|_2 & \alpha_{12}\|u_1\|_2 & \alpha_{13}\|u_1\|_2 \\ 0 & 0 & \|u_3\|_2 \end{bmatrix} \\ &= \left[\begin{array}{c|c} & \\ w_1 & w_3 \\ & \end{array}\right] \begin{bmatrix} \|u_1\|_2 & \alpha_{12}\|u_1\|_2 & \alpha_{13}\|u_1\|_2 \\ 0 & 0 & \|u_3\|_2 \end{bmatrix} \\ &= \underbrace{\begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \sqrt{3} & 2\sqrt{3} & \sqrt{3} \\ 0 & 0 & \sqrt{6} \end{bmatrix}}_{R}. \end{aligned}$$

With the Gram-Schmidt process and QR Factorization well-understood, we can discuss the formulation and solution of finite-dimensional least squares problems, which can be solved efficiently using this factorization. This will occur once linear operators are introduced in the next chapter.

4.9 Properties of Hilbert Spaces

Within this section, we use the notion of orthogonality to discuss two interesting geometric properties - the orthogonal projection of a vector onto a subspace and the orthogonal complement of a set. Additionally, we will introduce the direct sum of two sets and show that any Hilbert space can be decomposed into the direct sum of a closed subspace and its orthogonal complement.

Definition 4.26. Let $\mathcal{V}$ be a Hilbert space with $M \subseteq \mathcal{V}$. Then the **orthogonal complement of** M is defined by

$$M^{\perp} = \{v \in \mathcal{V} : \langle v, w \rangle = 0 \text{ for all } w \in M\}.$$

Example 32. Consider the plane

$$P = \{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0\}.$$

Then, $P^{\perp}$ is the set of all vectors that are orthogonal to this plane, which as you might be aware, form a line. In particular, these points lie along the normal vector of the plane, namely

$$P^{\perp} = \{x \in \mathbb{R}^3 : x = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \text{ for some } \alpha \in \mathbb{R}\}.$$

Notice that $P^{\perp}$ must contain 0, and thus translated lines that are parallel to the normal vector are not contained within the orthogonal complement.

Further, we note that within the definition of orthogonal complement M need only be a subset and not itself a subspace. Regardless, the next result shows that $M^{\perp}$ is always a subspace, even if M is not.

Theorem 4.23. For any $M \subseteq \mathcal{V}$, the orthogonal complement $M^{\perp}$ is a subspace of $\mathcal{V}$.

Proof. Certainly, $0 \in M^{\perp}$ because by Lemma 4.15

$$\langle 0, w \rangle = 0$$

for every $w \in M$. Additionally, if $u_1, u_2 \in M^{\perp}$ then

$$\langle u_1, w \rangle = \langle u_2, w \rangle = 0$$

for every $w \in M$. Therefore, we have for every $w \in M$

$$\langle u_1 + u_2, w \rangle = \langle u_1, w \rangle + \langle u_2, w \rangle = 0,$$

which implies $u_1 + u_2 \in M^{\perp}$. The closure of $M^{\perp}$ under scalar multiplication is similarly justified, and the proof is complete. □

Example 33. For any vector space $\mathcal{V}$ with an inner product, the orthogonal complement of the entire space $\mathcal{V}$ is just

$$\mathcal{V}^\perp = \{0\}.$$

Therefore, the only vector that is orthogonal to every vector in $\mathcal{V}$ is the zero vector. Hence, if a specific element $u \in \mathcal{V}$ satisfies

$$\langle u, v\rangle = 0$$

for every $v \in \mathcal{V}$, then it follows that $u = 0$.

Next, we will prove a crucial orthogonality result that will be needed in the proof of the Fundamental Theorem of Linear Algebra (Section 5.3).

Theorem 4.24 (Projection Theorem)**.** Let $\mathcal{V}$ be a Hilbert space and $M \subseteq \mathcal{V}$ be a closed subspace. Then, for any $v \in \mathcal{V}$ there exists a unique $v^* \in M$ such that

$$\delta := \|v - v^*\| = \inf_{u \in M} \|v - u\|.$$

Moreover, the difference is orthogonal to M, meaning $v - v^* \in M^\perp$.

Said another way, for any element v of the Hilbert space, there is a minimal distance from v to any element of the subspace M. In this way, we define δ to be the distance between the vector v and the subspace M.

Proof. We first prove the existence of $v^* \in M$. Let $v \in \mathcal{V}$ be given and define

$$\delta := \inf_{u \in M} \|v - u\|.$$

We must be a bit careful here because $\delta = 0$ or $\delta = \infty$ are both technically possible. By definition of the infimum [10], there is a sequence $u_n \in M$ such that

$$\delta_n = \|v - u_n\|$$

satisfies $\delta_n \to \delta$ as $n \to \infty$. The plan is to show that u_n is Cauchy, which will imply its convergence to a limit in M that we will define to be v^*. Let $w_n = v - u_n$ for every $n \in \mathbb{N}$ so that $\|w_n\| = \delta_n$ and

$$\|w_n + w_m\| = \|u_n + u_m - 2v\| = 2\left\|\frac{1}{2}(u_n + u_m) - v\right\| \geq 2\delta$$

because $\frac{1}{2}(u_n + u_m) \in M$. Furthermore, we find

$$\begin{aligned}
\|u_n - u_m\|^2 &= \|w_n - w_m\|^2 \\
&= \langle w_n - w_m, w_n - w_m\rangle \\
&= \|w_n\|^2 + \|w_m\|^2 - \langle w_n, w_m\rangle - \langle w_m, w_n\rangle \\
&= 2\|w_n\|^2 + 2\|w_m\|^2 - \|w_n\|^2 - \|w_m\|^2 - \langle w_n, w_m\rangle - \langle w_m, w_n\rangle \\
&= 2\left(\|w_n\|^2 + \|w_m\|^2\right) - \|w_n + w_m\|^2 \\
&\leq 2\left(\delta_n^2 + \delta_m^2\right) - (2\delta)^2 \\
&= 2\left(\delta_n^2 - \delta^2\right) + 2\left(\delta_m^2 - \delta^2\right).
\end{aligned}$$

Letting $m, n \to \infty$, we see that $\|u_n - u_m\|^2 \to 0$ because $\delta_n \to \delta$ and $\delta_m \to \delta$. Therefore, u_n is a Cauchy sequence in M. Because M is a closed subspace of $\mathcal{V}$, Theorem 4.18 implies that it is also complete, and hence u_n converges to a limit, denoted by $v^* \in M$. Because $v^* \in M$, we see that $\|v - v^*\| \geq \delta$ and

$$\|v - v^*\| \leq \|v - u_n\| + \|u_n - v^*\| = \delta_n + \|u_n - v^*\| \to \delta$$

as $n \to \infty$. Thus, $\|v - v^*\| \leq \delta$, and we conclude that $\|v - v^*\| = \delta$, so that the infimum is attained at v^*.

Now, to prove uniqueness we first assume that $v_1^* \in M$ and $v_2^* \in M$ satisfy

$$\|v - v_1^*\| = \delta \qquad \text{and} \qquad \|v - v_2^*\| = \delta.$$

Adding and subtracting as in the above calculation, we find

$$\begin{aligned} \|v_1^* - v_2^*\|^2 &= \|(v_1^* - v) - (v_2^* - v)\|^2 \\ &= 2\|v_1^* - v\|^2 + 2\|v_2^* - v\|^2 - \|(v_1^* - v) + (v_2^* - v)\|^2 \\ &= 2\delta^2 + 2\delta^2 - 4\left\|\frac{1}{2}(v_1^* + v_2^*) - v\right\|^2. \end{aligned}$$

Because M is a subspace, we see that $\frac{1}{2}(v_1^* + v_2^*) \in M$ and thus

$$\left\|\frac{1}{2}(v_1^* + v_2^*) - v\right\| \geq \delta.$$

Using this in the above series of equalities, we find

$$\|v_1^* - v_2^*\|^2 \leq 0,$$

and since this quantity must be nonnegative, we have $\|v_1^* - v_2^*\|^2 = 0$. Hence, $v_1^* = v_2^*$ and uniqueness is established.

Finally, we prove $v - v^* \in M^\perp$. Let $w = v - v^*$ so that $\|w\| = \delta$. Assume that $w \notin M^\perp$, then there exists $z \in M$ such that

$$\langle w, z \rangle = \alpha \neq 0.$$

Furthermore, $z \neq 0$ as this would imply $\langle w, z \rangle = 0$. Then, for any $\beta \in \mathbb{K}$, we find

$$\begin{aligned} \|w - \beta z\|^2 &= \langle w - \beta z, w - \beta z \rangle \\ &= \|w\|^2 - \beta\langle w, z\rangle - \bar{\beta}\overline{\langle w, z\rangle} + |\beta|^2\|z\|^2 \\ &= \|w\|^2 - \beta\alpha - \bar{\beta}\bar{\alpha} + |\beta|^2\|z\|^2 \\ &= \|w\|^2 - \bar{\beta}\bar{\alpha} - \beta\left(\alpha - \bar{\beta}\|z\|^2\right). \end{aligned}$$

Now, choosing $\bar{\beta} = \frac{\alpha}{\|z\|^2}$ so as to eliminate the final term on the right side, we find

$$\|w - \beta z\|^2 = \|w\|^2 - \frac{|\alpha|^2}{\|z\|^2} < \|w\|^2 = \delta^2,$$

which implies

$$\|w - \beta z\| < \delta.$$

However, since $w - \beta z = v - (v^* + \beta z)$ and $v^* + \beta z \in M$, we see that

$$\|w - \beta z\| \geq \inf_{u \in M} \|v - u\| = \delta.$$

This contradicts our original assumption $w \notin M^\perp$, thereby implying $w \in M^\perp$ and completing the proof. □

Comment. In view of this theorem, it is customary to refer to $v^* \in M$ as the **orthogonal projection** of the vector v onto the subspace M, as it is the unique element of M with $v - v^* \in M^\perp$. Of course, this is an idea that we've likely seen for vectors in $\mathbb{R}^n$ in a Linear Algebra course, though v^* is typically written as $\text{proj}_M v$ in that context.

Next, we define the direct sum of subspaces of a vector space.

Definition 4.27. Let $A, B \subseteq \mathcal{V}$ be subspaces of a vector space $\mathcal{V}$. Then, we say A and B form a **direct sum** for $\mathcal{V}$, written $A \bigoplus B = \mathcal{V}$, if

1. $A \cap B = \{0\}$
2. the set
$$A + B = \{a + b : a \in A, b \in B\}$$
satisfies $A + B = \mathcal{V}$.

Example 34. Here are two examples of direct sums - notice that each arises from a previously mentioned orthogonal complement.

1. Let $A = \mathcal{V}$ and $B = \{0\}$. Then,
$$A + B = \mathcal{V} \qquad \text{and} \qquad A \cap B = \{0\}.$$
Thus, $A \bigoplus B = \mathcal{V}$.
2. Let A be the plane
$$A = \{x \in \mathbb{R}^3 : x_1 + x_2 + x_3 = 0\}$$
and B be the line
$$B = \{x \in \mathbb{R}^3 : x = \alpha \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ for some } \alpha \in \mathbb{R}\}.$$
Then, $A \cap B = \{0\}$ because any vector $x \in B$ must satisfy
$$x_1 + x_2 + x_3 = 3\alpha$$
for some $\alpha \in \mathbb{R}$ and thus can only belong to A if $\alpha = 0$. Additionally, computing the parametric representation of A (i.e., the solution space of the equation of the plane), we find $x \in A$ implies
$$x = s \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix} + t \begin{bmatrix} 1 \\ -1 \\ 0 \end{bmatrix}$$
for some $s, t \in \mathbb{R}$. Since these two vectors, when combined with any nonzero element of B, form a basis for $\mathbb{R}^3$, we can decompose any element into a linear combination of these vectors. Thus, if $x \in \mathbb{R}^3$, then $x = a + b$ where $a \in A$ and $b \in B$. So $A \bigoplus B = \mathbb{R}^3$.

Now, we can connect the notions of orthogonal complement and direct sum to show that any Hilbert space can be decomposed into a direct sum of any closed subspace and its orthogonal complement.

Theorem 4.25 (Decomposition Theorem)**.** Let $\mathcal{V}$ be a Hilbert space and $M \subseteq \mathcal{V}$ be a closed subspace. Then, we may decompose $\mathcal{V}$ as

$$\mathcal{V} = M \bigoplus M^{\perp}.$$

Proof. By Theorem 4.24, for any $v \in \mathcal{V}$, there is a unique $v^* \in M$ (with minimal distance to v) such that

$$w = v - v^* \in M^{\perp}.$$

Hence, any $v \in \mathcal{V}$ can be written exactly as

$$v = \underbrace{v^*}_{\in M} + \underbrace{v - v^*}_{\in M^{\perp}},$$

which shows

$$\mathcal{V} = M + M^{\perp}.$$

Finally, $0 \in M$, as it's a subspace, and $0 \in M^{\perp}$, which means $0 \in M \cap M^{\perp}$. Contrastingly, if $u \in M$ and $u \in M^{\perp}$, then $u \perp u$, which means

$$\|u\|^2 = \langle u, u \rangle = 0,$$

and $u = 0$. This means that only 0 can be in both subspaces. Combining these statements yields

$$M \cap M^{\perp} = \{0\},$$

and thus

$$\mathcal{V} = M \bigoplus M^{\perp}.$$

□

Theorem 4.26. Let $\mathcal{V}$ be an inner product space, and $M \subseteq \mathcal{V}$ be a subspace. Then, we have

1. $M \subseteq M^{\perp\perp}$
2. If $\mathcal{V}$ is a Hilbert space and M is closed, then $M = M^{\perp\perp}$.

Proof. To prove the first portion, let $v \in M$ be given. Then, for any $w \in M^{\perp}$ we have $\langle v, w \rangle = 0$. So v is orthogonal to every element of $M^{\perp}$, which means $v \in (M^{\perp})^{\perp} = M^{\perp\perp}$. Thus, $M \subseteq M^{\perp\perp}$.

To prove the second consequence, we merely need to show $M^{\perp\perp} \subseteq M$. Let $v \in M^{\perp\perp} \subseteq \mathcal{V}$ be given. Then, by Theorem 4.25, there is $v^* \in M$ and $w \in M^{\perp}$ such that

$$v = v^* + w.$$

Now, the first conclusion of the theorem ($M \subseteq M^{\perp\perp}$) further implies $v^* \in M^{\perp\perp}$, and since $M^{\perp\perp}$ is a subspace of $\mathcal{V}$, we find $w = v - v^* \in M^{\perp\perp}$. However, $w \in M^{\perp}$, as well, so that

$$\|w\|^2 = \langle w, w \rangle = 0.$$

Of course, this implies $w = 0$ and $v = v^* \in M$. Finally, because $v \in M^{\perp\perp}$ was arbitrary, we find $M^{\perp\perp} \subseteq M$. □

Comment. There exist infinite-dimensional inner product spaces $\mathcal{V}$ in which subspaces $M \subseteq \mathcal{V}$ satisfy

$$M^{\perp\perp} \neq M.$$

In general, the **closure** of the subspace results from the double orthogonal complement, i.e.

$$M^{\perp\perp} = \overline{M},$$

but if M is closed as in Theorem 4.26, then $\overline{M} = M$.

The need for the closure occurs for the following reason. Consider a sequence $v_n \in M$ for all $n \in \mathbb{N}$ satisfying $v_n \to v$ where $v \in \mathcal{V}$ but $v \notin M$. Indeed, this can occur exactly when M is not closed. Then, for any $w \in M^{\perp}$, we have

$$\langle v_n, w \rangle = 0$$

for all $n \in \mathbb{N}$. Taking the limit as $n \to \infty$ of both sides and using the continuity of the inner product then gives

$$0 = \lim_{n\to\infty} \langle v_n, w \rangle = \langle \lim_{n\to\infty} v_n, w \rangle = \langle v, w \rangle.$$

Thus, $v \perp w$ for any $w \in M^{\perp}$, which means $v \in M^{\perp\perp}$, but $v \notin M$. With this, we see that limit points of sequences in M may be in $M^{\perp\perp}$ when M is not closed.

With the fundamental notions of vector spaces, and in particular Hilbert spaces, well understood. We will turn out attention to the study of linear operators defined on such spaces in the next chapter.

Exercises - Vector Spaces

Problem 4.1. Let $\mathcal{V}$ be a given set and $f : \mathbb{R} \to \mathcal{V}$ be a one-to-one and onto function. For every $u, v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, define the sum and scalar product operations on $\mathcal{V}$ by

$$u \oplus v = f\left(f^{-1}(u) + f^{-1}(v)\right) \qquad \text{and} \qquad \alpha \odot v = f\left(\alpha f^{-1}(v)\right).$$

Prove that $\mathcal{V}$ (defined over $\mathbb{R}$) is a vector space.

Problem 4.2. Let $\mathcal{V} = \mathbb{R}^+$ and for every $u, v \in \mathcal{V}$ and $\alpha \in \mathbb{R}$, define the sum and scalar product operations on $\mathcal{V}$ by

$$u \oplus v = uv \qquad \text{and} \qquad \alpha \odot v = v^\alpha.$$

Prove that $\mathcal{V}$ is a vector space.

Problem 4.3. Let $\mathbb{C}^p(\mathbb{R})$ be the vector space of complex-valued p-tuples defined over the field of real numbers, and $\mathbb{C}^p(\mathbb{C})$ be the vector space of complex-valued p-tuples defined over the field of complex numbers. Further, let

$$T_n = \left\{ \sum_{k=1}^{n} a_k \sin(k\pi x) : a_k \in \mathbb{R} \ \text{ for every } k = 1, ..., n \right\}$$

be the vector space of sinusoidal trigonometric polynomials of degree at most n and

$$C(0,1) = \{f : (0,1) \to \mathbb{R} \ \text{ such that } f \text{ is continuous}\}$$

be the vector space of continuous functions on $(0,1)$, both of which are defined over $\mathbb{R}$. For each of the following pairs of sets A and B, determine whether or not A is a subspace of B and justify your answer by providing either a proof or a counterexample.

(a) $A = \mathbb{R}^p$ and $B = \mathbb{C}^p(\mathbb{R})$.

(b) $A = \mathbb{R}^p$ and $B = \mathbb{C}^p(\mathbb{C})$

(c) $A = T_n$ for a given $n \in \mathbb{N}$ and $B = C(0,1)$.

(d) $A = \bigcup_{n \in \mathbb{N}} T_n$ and $B = C(0,1)$. **(Extra Credit)**

Problem 4.4. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces, let $M \subset \mathcal{V}$ be a subspace, and assume the transformation $T : \mathcal{V} \to \mathcal{W}$ satisfies

$$T(\alpha u + \beta v) = \alpha T(u) + \beta T(v) \tag{4.8}$$

for all $\alpha, \beta \in \mathbb{K}$ and $u, v \in \mathcal{V}$. Finally, define the set

$$T(M) := \{T(u) : u \in M\}.$$

(a) Show that $T(M)$ is a subspace of $\mathcal{W}$.

(b) Let $n \in \mathbb{N}$ be given and assume that the subset $S = \{v_1, ..., v_n\}$ spans M. Show that $T(S) = \{T(v_1), ..., T(v_n)\}$ spans $T(M)$.

Problem 4.5. Let $\mathcal{V}$ be a vector space with $v_1, ..., v_n \in \mathcal{V}$ for some $n \in \mathbb{N}$ and denote $S = \{v_1, .., v_n\}$.

(a) Prove that if S is linearly dependent and $v_{n+1}, ..., v_m \in \mathcal{V}$ for some $m \in \mathbb{N}$ with $m > n$, then the set $S \bigcup \{v_{n+1}, ..., v_m\}$ is also linearly dependent.

(b) Prove that if S is linearly independent then any non-empty subset of S is also linearly independent.

Hint: Proof by contradiction may be useful.

Problem 4.6. Let $u \in \mathbb{R}^p \setminus \{0\}$ and $v \in \mathbb{R}^q \setminus \{0\}$ be given and define $A = uv^T \in \mathbb{R}^{p \times q}$. Show that $\{u\}$ is a basis for

$$\text{Col}(A) = \{Ax : x \in \mathbb{R}^q\}$$

and determine the rank of A.

Problem 4.7. Let $\mathcal{V}$ be a vector space, with nested subspaces $\mathcal{V}_0 \subseteq \mathcal{V}_1 \subseteq \mathcal{V}$ satisfying $\dim(\mathcal{V}_0) = \dim(\mathcal{V}_1) < \infty$. Prove that $\mathcal{V}_0 = \mathcal{V}_1$.
Hint: Proof by contradiction may be useful.

Problem 4.8. Define the vector space $\ell^\infty(\mathbb{R})$ to be the set of all bounded infinite sequences of real numbers $\{a_n\}_{n=1}^\infty$. Prove that $\dim(\ell^\infty(\mathbb{R})) = \infty$.
Hint: Proof by contradiction may be useful.

Problem 4.9. Prove (via inequalities) that the norms $\|\cdot\|_1$, $\|\cdot\|_2$, and $\|\cdot\|_\infty$ defined on $\mathbb{R}^p$ are equivalent.

Problem 4.10. A subset A of a vector space $\mathcal{V}$ is **convex** if for any $u, v \in A$, the line segment

$$H = \{\alpha u + (1 - \alpha)v : 0 \leq \alpha \leq 1\}$$

is a subset of A.

(a) Show that for any normed space $\mathcal{V}$, the closed unit ball

$$B(0,1) = \{v \in \mathcal{V} : \|v\| \leq 1\}$$

is convex.

(b) Show that the function $\phi : \mathbb{R}^2 \to \mathbb{R}$ defined by

$$\phi(v) = \left(\sqrt{|v_1|} + \sqrt{|v_2|}\right)^2$$

is not a norm on $\mathbb{R}^2$.

(c) Sketch the curve $\phi(v) = 1$.

Problem 4.11. Let $p, q \in \mathbb{N}$ be given.

(a) Let A be a real, symmetric $p \times p$ matrix satisfying $x^T Ax > 0$ for every $x \in \mathbb{R}^p \setminus \{0\}$. Prove that the function

$$\langle x, y \rangle_A := x^T Ay$$

defined for any $x, y \in \mathbb{R}^p$ is an inner product on $\mathbb{R}^p$.

(b) Prove that the function

$$\|A\|_F := \sqrt{\text{tr}(A^T A)}$$

defined for any $A \in \mathbb{R}^{p \times q}$ is a norm on $\mathbb{R}^{p \times q}$.
Hint: Use a formula for $\|A\|_F$ in terms of the columns or entries of A.

Problem 4.12. Let $\{v_n\}_{n=1}^{\infty}$ be a sequence of vectors in a Hilbert space $\mathcal{V}$ with

$$\|v_n\| \to \|v\| \quad \text{and} \quad \langle v_n, v \rangle \to \|v\|^2.$$

Show that v_n converges to v in $\mathcal{V}$.

Problem 4.13. Let $n \in \mathbb{N}$ be given and assume $S = \{v_1, ..., v_n\} \subseteq \mathcal{V}$ is orthogonal with $0 \notin S$. Show that S is linearly independent.

Problem 4.14. Let $\mathcal{V}$ be a Hilbert space over $\mathbb{K}$. Show that if $S = \{v_1, ..., v_q\} \subset \mathcal{V}$ with $q \in \mathbb{N}$ is orthonormal, then for every $v \in \mathcal{V}$ we have

$$\sum_{k=1}^{q} |\alpha_k|^2 \leq \|v\|^2 \quad \text{where} \quad \alpha_k = \langle v_k, v \rangle.$$

Hint: Consider $w = \sum_{k=1}^{q} \alpha_k v_k$ and $\langle v, w \rangle$.

Problem 4.15. Consider the Hilbert space $\mathcal{V} = L^2(-1, 1)$ defined over $\mathbb{R}$ and endowed with the inner product

$$\langle f, g \rangle_2 := \int_{-1}^{1} f(x)g(x)\, dx$$

for every $f, g \in \mathcal{V}$. Let $S = \{1, x, x^2\} \subset \mathcal{V}$.

(a) Show that S is not an orthogonal set with respect to this inner product.

(b) Construct an orthonormal basis for span(S).

Remark: You are constructing the first three terms of a basis known as the *orthonormal Legendre Polynomials*, which is generated by using $S = \{x^n : n \in \mathbb{N}_0\}$ and continuing this process. The resulting polynomials form an orthonormal basis for $L^2(-1, 1)$.

Problem 4.16. Let $\mathcal{V}$ be a Hilbert space. Prove that if $u, v \in \mathcal{V}$ with $u \perp v$ then

$$\|u + v\|^2 = \|u\|^2 + \|v\|^2.$$

This is known as the Pythagorean Theorem.

Problem 4.17. Let A be a real $p \times q$ matrix satisfying $A = QR$, where Q is a real $p \times k$ matrix with orthonormal columns and R is a real $k \times q$ matrix with rank$(R) = k$.

(a) Show that R has a right inverse; that is, show that there exists a real $q \times k$ matrix X such that $RX = I_k$.

(b) Show that the columns of Q form an orthonormal basis for Col(A).

Chapter 5

Linear Operators on Vector Spaces

As before, we assume throughout that all vector spaces are either real or complex, i.e. defined over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Now that we have introduced a variety of subcategories of vector spaces, we will generally assume that $\mathcal{V}$ and $\mathcal{W}$ are Hilbert spaces. However, when a weaker assumption (e.g., a Banach space or general vector space without a norm) can be utilized, we will state this explicitly.

5.1 Introduction and Definitions

We begin by defining a linear operator (or mapping) on a vector space.

Definition 5.1. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces. A **linear operator** $T : \mathcal{V} \to \mathcal{W}$ is a function that assigns to each $v \in \mathcal{V}$ a unique $T(v) \in \mathcal{W}$ such that

1. For every $u, v \in \mathcal{V}$, we have

$$T(u + v) = T(u) + T(v)$$

2. For every $\alpha \in \mathbb{K}$, $v \in \mathcal{V}$, we have

$$T(\alpha v) = \alpha T(v).$$

Comment. To show that a given operator T is linear, it suffices to prove

$$T(\alpha u + \beta v) = \alpha T(u) + \beta T(v)$$

for every $\alpha, \beta \in \mathbb{K}$ and $u, v \in \mathcal{V}$.

Theorem 5.1. Assume $T : \mathcal{V} \to \mathcal{W}$ is linear. Then, $T(0) = 0$.

Proof. We merely use the aforementioned properties of linearity to compute for any $v \in \mathcal{V}$

$$T(0) = T(0 \cdot v) = 0 \cdot T(v) = 0,$$

which completes the proof. Additionally, we note that the last term in this string of equalities (on the right) represents the zero vector in $\mathcal{W}$, while the zero within the argument of the first term (on the left) represents the zero vector in $\mathcal{V}$, and these need not be the same object. □

Though linearity of an operator may seem like a fairly restrictive assumption with which to begin, standard methods for the study of nonlinear operators have been both elusive and the subject of much ongoing mathematical research. Additionally, we will often use another property of the operators we study, namely boundedness.

Definition 5.2. Let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces. A linear operator $T : \mathcal{V} \to \mathcal{W}$ is **bounded** if there is $C > 0$ such that for all $v \in \mathcal{V}$

$$\|T(v)\|_{\mathcal{W}} \leq C\|v\|_{\mathcal{V}}.$$

If T is not bounded, then we say it is **unbounded**.

Comment. Though it is left as an exercise (cf. Problem 5.7), one remarkable result concerning bounded linear operators is that they are equivalent to continuous linear operators. In particular, this means that if $v_n \to v$ in $\mathcal{V}$, then $T(v_n) \to T(v)$ in $\mathcal{W}$ for any bounded linear operator $T : \mathcal{V} \to \mathcal{W}$. This property is crucial on infinite-dimensional Banach spaces since the continuity of T guarantees that the linearity property can be extended to infinite sums, for instance

$$T\left(\sum_{k=1}^{\infty} \alpha_k v_k\right) = T\left(\lim_{N\to\infty} \sum_{k=1}^{N} \alpha_k v_k\right) = \lim_{N\to\infty} T\left(\sum_{k=1}^{N} \alpha_k v_k\right) = \sum_{k=1}^{\infty} \alpha_k T(v_k)$$

and we will certainly use this property in the future.

Before continuing, we need some intuition and examples of these objects.

Example 35. There are many prominent examples of linear operators, including matrices, integrals, and derivatives.

1. Let $A \in \mathbb{R}^{p\times q}$ be given. Then, $T : \mathbb{R}^q \to \mathbb{R}^p$ defined by

   $$T(v) = Av$$

   for every $v \in \mathbb{R}^q$ is a bounded, linear operator.

2. Let $a, b \in \mathbb{R}$ with $a < b$ be given and define $L : C[a, b] \to \mathbb{R}$ by

   $$L[f] = \int_a^b f(x)\, dx.$$

   Then, L is a bounded, linear operator.

3. Let $a, b \in \mathbb{R}$ with $a < b$ be given and define $D : C^\infty(a, b) \to C^\infty(a, b)$ by

   $$D[f] = f'(x)$$

   Then, D is an unbounded, linear operator. Indeed, the sequence of functions $f_n(x) = \sin(nx) \in C^\infty(a, b)$ satisfies $|f_n(x)| \leq 1$ for every $n \in \mathbb{N}, x \in (a, b)$, but their derivatives $D[f_n] = f_n'(x) = n\cos(nx)$ are unbounded in $C^\infty(a, b)$.

Definition 5.3. Let $T : \mathcal{V} \to \mathcal{W}$ be linear. Then, we define the **range** (or **image)** of T, denoted $\mathrm{R}(T)$, and the **kernel** of T, denoted $\mathrm{Ker}(T)$, by

$$\mathrm{R}(T) = \{T(v) : v \in \mathcal{V}\}$$
$$\mathrm{Ker}(T) = \{v \in \mathcal{V} : T(v) = 0\}.$$

Of course, in the case that $T(v) = Av$ for some $A \in \mathbb{R}^{p\times q}$, we see that $\mathrm{R}(T) \subseteq \mathbb{R}^p$ and $\mathrm{R}(T) = \mathrm{Col}(A)$, while $\mathrm{Ker}(T) \subseteq \mathbb{R}^q$ and $\mathrm{Ker}(T) = \mathrm{Nul}(A)$. So, these should be familiar sets, and just like $\mathrm{Col}(A)$ and $\mathrm{Nul}(A)$ they should also be subspaces.

Theorem 5.2. Let $T : \mathcal{V} \to \mathcal{W}$ be linear. Then, $\mathrm{R}(T)$ is a subspace of $\mathcal{W}$ and $\mathrm{Ker}(T)$ is a subspace of $\mathcal{V}$.

Proof. To show that either set is a subspace, we must show that it is a subset that contains the zero vector and is closed under both addition and scalar multiplication. To this end, we note that Theorem 5.1 implies $T(0) = 0$. Hence, $0 \in \mathrm{Ker}(T) \subseteq \mathcal{V}$ and $0 \in \mathrm{R}(T) \subseteq \mathcal{W}$ since $0 \in \mathcal{V}$. Next, we let $w, z \in \mathrm{R}(T)$ be given. Then, there are $u, v \in \mathcal{V}$ such that

$$T(u) = w \qquad \text{and} \qquad T(v) = z.$$

Hence, letting $x = u + v$, we find by linearity of T

$$w + z = T(u) + T(v) = T(u + v) = T(x) \in \mathrm{R}(T).$$

Thus, $w + z \in \mathrm{R}(T)$ and $\mathrm{R}(T)$ is closed under addition. Similar computations can be performed to show the closure under scalar multiplication, and analogously these properties can be shown for $\mathrm{Ker}(T)$. □

With basic terminology and minor results out of the way, we turn our attention to a much more substantial theorem involving these subspaces.

Theorem 5.3 (Generalized Rank-Nullity)**.** Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces and $T : \mathcal{V} \to \mathcal{W}$ be linear with $\dim(\mathcal{V}) = q \in \mathbb{N}$. Then,

1. We have
$$\dim(\mathrm{Ker}(T)) \leq q \qquad \text{and} \qquad \dim(\mathrm{R}(T)) \leq q.$$
2. If there is $B_1 = \{v_1, ..., v_k\} \subseteq \mathcal{V}$ such that the set $\{T(v_1), ..., T(v_k)\} \subseteq \mathcal{W}$ is a basis for $\mathrm{R}(T)$ and $B_2 = \{u_1, ..., u_n\} \subseteq \mathcal{V}$ is a basis for $\mathrm{Ker}(T)$, then
$$B = B_1 \cup B_2 = \{v_1, ..., v_k, u_1, ..., u_n\}$$
is a basis for $\mathcal{V}$.
3. We have
$$\dim(\mathcal{V}) = \dim(\mathrm{Ker}(T)) + \dim(\mathrm{R}(T)).$$

Proof. To prove the first conclusion, we begin by noting that $\mathrm{Ker}(T)$ is a subspace of $\mathcal{V}$ by Theorem 5.2. Then, by Theorem 4.9, the dimension of any subspace must be no greater than the dimension of the space itself, and we see that

$$\dim(\mathrm{Ker}(T)) \leq \dim(\mathcal{V}) = q.$$

Next, let $w \in \mathrm{R}(T)$ be given so that there is some $v \in \mathcal{V}$ with $T(v) = w$. Since $\dim(\mathcal{V}) = q$, $\mathcal{V}$ has a basis of q elements, say $B_{\mathcal{V}} = \{v_1, ..., v_q\} \subseteq \mathcal{V}$. Certainly, we can express v in terms of this basis; so there are $\alpha_j \in \mathbb{K}$ for $j = 1, ..., q$ with $v = \sum_{j=1}^{q} \alpha_j v_j$. Therefore, we find

$$w = T(v) = T\left(\sum_{j=1}^{q} \alpha_j v_j\right) = \sum_{j=1}^{q} \alpha_j T(v_j).$$

Thus, w can be expressed as a linear combination of the image of basis elements. Since $w \in \mathrm{R}(T)$ was arbitrary, we see that the set $\{T(v_1), ..., T(v_q)\}$ spans $\mathrm{R}(T)$. By Lemma 4.7, it follows that any linearly independent subset of $\mathrm{R}(T)$ must contain at most q vectors. Therefore, any basis of $\mathrm{R}(T)$ must contain at most q vectors, and by definition $\dim(\mathrm{R}(T)) \leq q$.

The second conclusion of the theorem will be a homework problem (cf. Problem 5.1). We provide a brief sketch for one component of the result:

We wish to show that B spans $\mathcal{V}$. Let $v \in \mathcal{V}$ be given. Define $w = T(v)$ so that $w \in R(T)$. Since $\{T(v_1), \ldots, T(v_k)\}$ is a basis for $R(T)$, there are $\alpha_1, \ldots, \alpha_k \in \mathbb{K}$ such that

$$w = \sum_{j=1}^{k} \alpha_j T(v_j).$$

Said another way, this is equivalent to

$$T(v) = \sum_{j=1}^{k} \alpha_j T(v_j).$$

Therefore, using the linear properties of T, we find

$$T\left(v - \sum_{j=1}^{k} \alpha_j v_j\right) = T(v) - T\left(\sum_{j=1}^{k} \alpha_j v_j\right) = T(v) - \sum_{j=1}^{k} \alpha_j T(v_j) = 0.$$

Thus,

$$v - \sum_{j=1}^{k} \alpha_j v_j \in \mathrm{Ker}(T).$$

Since $B_2 = \{u_1, \ldots, u_n\}$ is a basis for $\mathrm{Ker}(T)$, there are $\beta_1, \ldots, \beta_n \in \mathbb{K}$ such that

$$v - \sum_{j=1}^{k} \alpha_j v_j = \sum_{\ell=1}^{n} \beta_\ell u_\ell.$$

Finally, consolidating the sums on the right side of the equation yields

$$v = \sum_{j=1}^{k} \alpha_j v_j + \sum_{\ell=1}^{n} \beta_\ell u_\ell,$$

which is just a linear combination of elements from B. Since $v \in \mathcal{V}$ was arbitrary, B spans $\mathcal{V}$. It remains to show that B is linearly independent.

Nicely, the last conclusion follows straightforwardly from the second. Indeed, if B_1 is a basis for $\mathrm{R}(T)$, then $\dim(\mathrm{R}(T)) = k$, while $\dim(\mathrm{Ker}(T)) = n$ if B_2 is a basis for $\mathrm{Ker}(T)$. Finally, because B is a basis for $\mathcal{V}$, we have

$$\dim(\mathcal{V}) = n + k = \dim(\mathrm{Ker}(T)) + \dim(\mathrm{R}(T)),$$

and the proof is complete. □

Comment. There are a number of remarks that accompany this result.

1. Notice that $\mathcal{W}$ does not need to be finite dimensional.

2. $\dim(\{0\}) = 0$ and $\{0\}$ is the only zero-dimensional subspace of $\mathcal{V}$.

3. If $\dim(\mathrm{R}(T)) = 0$, then $\mathrm{R}(T) = \{0\}$ and Theorem 5.3 implies $\dim(\mathrm{Ker}(T)) = \dim(\mathcal{V})$. Thus, it follows that $\mathrm{Ker}(T) = \mathcal{V}$.

4. If $\dim(\mathrm{Ker}(T)) = 0$, then $\mathrm{Ker}(T) = \{0\}$ and Theorem 5.3 implies $\dim(\mathrm{R}(T)) = \dim(\mathcal{V})$. Note, however, that $\mathrm{R}(T)$ is not a subset of $\mathcal{V}$, so $\mathrm{R}(T) \neq \mathcal{V}$ necessarily. Additionally, this does not imply $\mathrm{R}(T) = \mathcal{W}$ because $\mathcal{W}$ could be much larger; in fact it's possible that $\dim(\mathcal{W}) = \infty$.

5. If $\mathcal{V}$ is infinite-dimensional, then the generalized Rank-Nullity Theorem still holds, but with generalized cardinal arithmetic.

Example 36. Of course, in the case that $\mathcal{W}$ is finite-dimensional, this theorem is well-known from Linear Algebra. In particular, if $A \in \mathbb{R}^{p\times q}$ and $T : \mathbb{R}^q \to \mathbb{R}^p$ is defined by $T(v) = Av$ for every $v \in \mathbb{R}^q$, then T is linear and we have

$$\begin{aligned}\mathrm{Ker}(T) &= \{v \in \mathbb{R}^q : T(v) = 0\} = \{v \in \mathbb{R}^q : Av = 0\} = \mathrm{Nul}(A)\\ \mathrm{R}(T) &= \{T(v) : v \in \mathbb{R}^q\} = \{Av : v \in \mathbb{R}^q\} = \mathrm{Col}(A).\end{aligned}$$

Additionally, $\mathrm{rank}(A) = \dim(\mathrm{Col}(A))$ and therefore, Theorem 5.3 implies

$$q = \dim(\mathbb{R}^q) = \mathrm{rank}(A) + \dim(\mathrm{Nul}(A)),$$

which is the result that gives the Rank-Nullity theorem its name.

Next, we focus on how $\mathrm{R}(T)$ and $\mathrm{Ker}(T)$ influence our ability to solve the linear equation $T(v) = w$ and, more generally, the notion of invertibility of linear trasnformations.

Definition 5.4. Assume $T : \mathcal{V} \to \mathcal{W}$ is linear and let $w \in \mathcal{W}$ be given. We say the equation $T(v) = w$ is **solvable** if there is $u \in \mathcal{V}$ such that $T(u) = w$. In this case, u is called the **solution** of the equation $T(v) = w$.

Theorem 5.4. Let $T : \mathcal{V} \to \mathcal{W}$ be linear. Then,

1. Given $w \in \mathcal{W}$, the equation $T(v) = w$ is solvable if and only if $w \in \mathrm{R}(T)$.

2. The equation $T(v) = w$ is solvable for every $w \in \mathcal{W}$ if and only if $\mathrm{R}(T) = \mathcal{W}$ (i.e., T is onto - recall Definition 4.8).

Proof. The proof is fairly straightforward and left as an exercise. □

Theorem 5.5. Let $w \in \mathcal{W}$ be given and assume $T : \mathcal{V} \to \mathcal{W}$ is linear and $T(v) = w$ is solvable with solution $u \in \mathcal{V}$. Then, u is the unique solution of $T(v) = w$ if and only if $\mathrm{Ker}(T) = \{0\}$.

Proof. We first assume $u \in \mathcal{V}$ is the unique solution of $T(v) = w$, and note that $0 \in \mathrm{Ker}(T)$ so that we need only show that $\mathrm{Ker}(T) \subseteq \{0\}$. Let $z \in \mathrm{Ker}(T)$ be given, which ensures $T(z) = 0$. Then, we find

$$T(u + z) = T(u) + T(z) = w + 0 = w.$$

Therefore, $u+z$ is a solution of $T(v)=w$. However, since u was assumed to be the unique solution of this equation, we must have $u+z=u$. Subtracting u implies $z=0$. Since $z\in\text{Ker}(T)$ was arbitrary, we have shown that $\text{Ker}(T)\subseteq\{0\}$, and thus $\text{Ker}(T)=\{0\}$.

Next, assume $\text{Ker}(T)=\{0\}$, and let $z\in\mathcal{V}$ be another solution of $T(v)=w$. Then, we find

$$T(u-z)=T(u)-T(z)=w-w=0.$$

Thus, $u-z\in\text{Ker}(T)$, and since 0 is the only element of this subspace, this means $u-z=0$ or $u=z$. Since z was an arbitrary solution of the equation, we see that u must be the unique solution. □

Combining the two previous theorems provides the following simplified result.

Theorem 5.6. Let $T:\mathcal{V}\to\mathcal{W}$ be linear. Then,

1. Given $w\in\mathcal{W}$, the equation $T(v)=w$ has a unique solution if and only if $w\in\text{R}(T)$ and $\text{Ker}(T)=\{0\}$.
2. The equation $T(v)=w$ has a unique solution for every $w\in\mathcal{W}$ if and only if $\text{R}(T)=\mathcal{W}$ (i.e, T is onto) and $\text{Ker}(T)=\{0\}$.

This last portion of Theorem 5.6 is crucial to define the inverse of a linear operator.

Definition 5.5. Assume $T:\mathcal{V}\to\mathcal{W}$ is linear, onto, and $\text{Ker}(T)=\{0\}$. Under these assumptions, we say T is **invertible** and define its inverse operator $T^{-1}:\mathcal{W}\to\mathcal{V}$ by assigning to every $w\in\mathcal{W}$, the unique $v\in\mathcal{V}$ such that $T(v)=w$ to denote the vector $T^{-1}(w)$. Said another way,

$$T(v)=w\iff v=T^{-1}(w).$$

Example 37. Consider $A\in\mathbb{R}^{p\times q}$ and define $T:\mathbb{R}^q\to\mathbb{R}^p$ by $T(v)=Av$.

1. If we impose the condition that T is onto, then $R(T)=\text{Col}(A)=\mathbb{R}^p$. Thus, $\text{rank}(A)=p$, and the q columns of A span $\mathbb{R}^p$. Since $\mathbb{R}^p$ possesses a basis (and thus, a linearly independent set) with p elements, it follows from Lemma 4.7 that $p\leq q$.
2. If we impose the condition $\text{Ker}(T)=\{0\}$, then $Av=0$ has only the solution $v=0$, which means that the q columns of A are linearly independent elements of $\mathbb{R}^p$. Of course, $\mathbb{R}^p$ has a basis (and thus, a spanning set) containing p vectors. Hence, it follows from Lemma 4.7 that $q\leq p$.

Thus, these two conditions imply $p=q$. Furthermore, we found that the columns of A are linearly independent. Hence, the Invertible Matrix Theorem implies that A is invertible, and these are exactly the conditions necessary and sufficient to arrive at an invertible matrix.

Theorem 5.7. Assume $T:\mathcal{V}\to\mathcal{W}$ is linear and invertible. Then,

1. The operator $T^{-1}:\mathcal{W}\to\mathcal{V}$ is linear.

2. For every $v \in \mathcal{V}$ and $w \in \mathcal{W}$, we have

$$T^{-1}(T(v)) = v \qquad \text{and} \qquad T(T^{-1}(w)) = w.$$

Proof. To prove the first conclusion, we let $w, z \in \mathcal{W}$ be given and define $v = T^{-1}(w)$ and $u = T^{-1}(z)$. Then, $T(v) = w$ and $T(u) = z$. So,

$$w + z = T(v) + T(u) = T(v + u).$$

Since $v + u \in \mathcal{V}$, by definition of the inverse we have

$$T^{-1}(w + z) = v + u = T^{-1}(w) + T^{-1}(z).$$

Showing the scalar multiplication property is similar.

To prove the second assertion, we let $v \in \mathcal{V}$ be given and set $w = T(v)$. By definition, we may write $v = T^{-1}(w)$. With this, we find

$$T^{-1}(T(v)) = T^{-1}(w) = v$$

for any $v \in \mathcal{V}$, and a similar argument for $T(T^{-1}(w))$ holds. □

5.2 The Adjoint Operator

Our main goal in subsequent sections is to state and prove the Fundamental Theorem of Linear Algebra, which describes the fundamental relationships between four special subspaces induced by every linear operator. Prior to this, however, we need to define the adjoint of a linear operator - an essential object that functionally generalizes the transpose of a matrix.

Definition 5.6. Let $\mathcal{V}$ and $\mathcal{W}$ be Hilbert spaces with inner products $\langle \cdot, \cdot \rangle_{\mathcal{V}}$ and $\langle \cdot, \cdot \rangle_{\mathcal{W}}$, and let $T : \mathcal{V} \to \mathcal{W}$ be linear. Then, (if it exists) the **adjoint of** T, denoted $T^* : \mathcal{W} \to \mathcal{V}$, is a linear operator satisfying

$$\langle T(v), w \rangle_{\mathcal{W}} = \langle v, T^*(w) \rangle_{\mathcal{V}}$$

for every $v \in \mathcal{V}$ and $w \in \mathcal{W}$.

Comment. The adjoint operator is extremely useful in studying partial differential equations or performing sensitivity analysis of established scientific models. Here are a few other comments:

1. If T is bounded, then T^* is guaranteed to exist and is also bounded, though this may not hold for unbounded operators on Hilbert spaces. This fact follows from the Riesz Representation Theorem [10], which (though an amazing result) is outside the scope of the course.

2. If $\mathcal{V}$ and $\mathcal{W}$ are finite dimensional then every linear operator is bounded (cf. Problem 5.6); hence, T^* is guaranteed to exist, while this may not always hold in the infinite-dimensional framework.

3. Additionally, if the adjoint T^* is known to exist, then T^* also has an adjoint and it is T, meaning $T^{**} = T$.

Next, we show that the adjoint, whenever it exists, must be unique.

Theorem 5.8. Assume $T : \mathcal{V} \to \mathcal{W}$ and $S : \mathcal{W} \to \mathcal{V}$ are linear and satisfy

$$\langle T(v), w\rangle_{\mathcal{W}} = \langle v, S(w)\rangle_{\mathcal{V}}$$

for every $v \in \mathcal{V}$ and $w \in \mathcal{W}$. Then, S is the unique linear operator satisfying this relationship and $S = T^*$.

Proof. Of course, if S satisfies this property, we define $T^* = S$. To show uniqueness, we let $R : \mathcal{W} \to \mathcal{V}$ be linear and satisfy the adjoint property. Then, we have for every $v \in \mathcal{V}$ and $w \in \mathcal{W}$,

$$\langle T(v), w\rangle_{\mathcal{W}} = \langle v, S(w)\rangle_{\mathcal{V}} = \langle v, R(w)\rangle_{\mathcal{V}}.$$

Using the last equality, we subtract to find

$$\langle v, S(w) - R(w)\rangle_{\mathcal{V}} = 0$$

for every $v \in \mathcal{V}$ and $w \in \mathcal{W}$. Hence, for any $w \in \mathcal{W}$ we choose

$$v = S(w) - R(w) \in \mathcal{V}$$

so that the previous equality is just

$$0 = \langle v, v\rangle_{\mathcal{V}} = \|v\|_{\mathcal{V}}^2,$$

which implies $v = 0$. Thus, $S(w) - R(w) = 0$, and $R(w) = S(w)$ for every $w \in \mathcal{W}$, which completes the uniqueness proof. □

To gain some intuition about this abstract operator, we consider a few examples, both from finite- and infinite-dimensional spaces.

Example 38. Let $A \in \mathbb{R}^{p\times q}$ and define $T : \mathbb{R}^q \to \mathbb{R}^p$ by $T(v) = Av$ for every $v \in \mathbb{R}^q$. Then, because $\mathbb{R}^p$ is endowed with the inner product

$$\langle x, y\rangle_{\mathbb{R}^p} = x^T y,$$

and using Corollary 2.1, we have for every $x \in \mathbb{R}^q$ and $y \in \mathbb{R}^p$

$$\begin{aligned}\langle T(x), y\rangle_{\mathbb{R}^p} &= \langle Ax, y\rangle_{\mathbb{R}^p}\\ &= (Ax)^T y\\ &= x^T A^T y\\ &= \langle x, A^T y\rangle_{\mathbb{R}^q}\\ &= \langle x, T^*(y)\rangle_{\mathbb{R}^q}\end{aligned}$$

Hence, the adjoint of T is $T^* : \mathbb{R}^p \to \mathbb{R}^q$ defined for every $v \in \mathbb{R}^p$ by

$$T^*(v) = A^T v.$$

Example 39. If we change $\mathbb{R}^p$ to $\mathbb{C}^p$ and $\mathbb{R}^q$ to $\mathbb{C}^q$, then the associated adjoint of T is $T^* : \mathbb{C}^p \to \mathbb{C}^q$ defined by

$$T^*(v) = A^H v,$$

for every $v \in \mathbb{C}^p$.

Example 40. Recall the vector space

$$C^\infty[0,1] = \bigcap_{n=1}^{\infty} C^n[0,1]$$

to be the set of all functions f that possesses arbitrarily-many continuous derivatives, and define the subspace

$$C_0^\infty[0,1] = \{f \in C^\infty[0,1] : f^{(n)}(0) = f^{(n)}(1) = 0 \text{ for all } n \in \mathbb{N}_0\}$$

where $f^{(n)}$ represents the nth derivative of the function f. Next, define an inner product on $C_0^\infty[0,1]$ by

$$\langle f, g\rangle = \int_0^1 f(x)g(x)dx$$

for every $f, g, \in C_0^\infty[0,1]$. You can (and should) verify that this is, in fact, an inner product. Finally, define the linear operator $T : C_0^\infty[0,1] \to C_0^\infty[0,1]$ by

$$T[f] = \frac{d^2 f}{dx^2}.$$

Notice that the second derivative of an infinitely-differentiable function is again infinitely-differentiable and satisfies the same boundary conditions, thereby implying $T[f] \in C_0^\infty[0,1]$ for every $f \in C_0^\infty[0,1]$.

Then, what is T^*? Computing the inner product of $T[f]$ with g for any $f, g \in C_0^\infty[0,1]$, we find by integration by parts

$$\begin{aligned}
\langle T[f], g\rangle &= \int_0^1 f''(x)g(x)dx \\
&= -\int_0^1 f'(x)g'(x)dx + \cancelto{0}{f'(x)g(x)\big|_{x=0}^{1}} \\
&= \int_0^1 f(x)g''(x)dx \\
&= \langle f, T[g]\rangle.
\end{aligned}$$

Therefore, we see that $T^*[g] = T[g]$. Thus, $T^* = T$, in which case, we say that T is **self-adjoint**.

Next, we will use the adjoint operator to obtain one of the more interesting results concerning the geometry of Hilbert spaces.

5.3 The Fundamental Theorem of Linear Algebra

Now, we're able to state and prove the Fundamental Theorem. For additional information and geometric insight into this result, the world-renowned applied mathematician Gilbert Strang has written a very nice review [23].

Theorem 5.9 (Fundamental Theorem of Linear Algebra)**.** Let $\mathcal{V}$ and $\mathcal{W}$ be Hilbert spaces. Let $T : \mathcal{V} \to \mathcal{W}$ be a bounded linear operator, and denote its adjoint by $T^* : \mathcal{W} \to \mathcal{V}$. Then, we have

1. $\mathrm{Ker}(T) = \mathrm{R}(T^*)^\perp$
2. $\mathrm{Ker}(T^*) = \mathrm{R}(T)^\perp$

and if we further assume $\mathrm{R}(T)$ is closed, then

3. $\mathrm{R}(T) = \mathrm{Ker}(T^*)^\perp$
4. $\mathrm{R}(T^*) = \mathrm{Ker}(T)^\perp$.

Proof. First, notice that the existence of T^* is guaranteed by the boundedness of T. Now, we establish the first conclusion, and note that the remaining conclusions will follow from this. Recalling the definition of the kernel, range, adjoint, and orthogonal complement, we write

$$\begin{aligned}\mathrm{Ker}(T) &= \{v \in \mathcal{V} : T(v) = 0\} \\ \mathrm{R}(T^*) &= \{T^*(w) : w \in \mathcal{W}\}\end{aligned}$$

so that

$$\begin{aligned}\mathrm{R}(T^*)^\perp &= \{v \in \mathcal{V} : \langle v, T^*(w)\rangle_{\mathcal{V}} = 0 \ \text{ for all } w \in \mathcal{W}\} \\ &= \{v \in \mathcal{V} : \langle T(v), w\rangle_{\mathcal{W}} = 0 \ \text{ for all } w \in \mathcal{W}\}.\end{aligned}$$

We will show separately that $\mathrm{Ker}(T) \subseteq \mathrm{R}(T^*)^\perp$ and $\mathrm{R}(T^*)^\perp \subseteq \mathrm{Ker}(T)$ in order to establish the equality of these two sets.

First, let $v \in \mathrm{Ker}(T)$ be given so that $T(v) = 0$. Then, for any $w \in \mathcal{W}$

$$\langle v, T^*(w)\rangle_{\mathcal{V}} = \langle T(v), w\rangle_{\mathcal{W}} = \langle 0, w\rangle_{\mathcal{W}} = 0.$$

Thus, $v \in \mathrm{R}(T^*)^\perp$, and we have $\mathrm{Ker}(T) \subseteq \mathrm{R}(T^*)^\perp$.

Next, let $v \in \mathrm{R}(T^*)^\perp$ be given. Then, for any $w \in \mathcal{W}$, we find

$$\langle T(v), w\rangle_{\mathcal{W}} = \langle v, T^*(w)\rangle_{\mathcal{V}} = 0.$$

Since this holds for every $w \in \mathcal{W}$, choose $w = T(v) \in \mathcal{W}$. Hence, we find

$$\|T(v)\|_{\mathcal{W}}^2 = \langle T(v), T(v)\rangle_{\mathcal{W}} = 0$$

and so $T(v) = 0$. Thus, $v \in \mathrm{Ker}(T)$, and we have shown $\mathrm{R}(T^*)^\perp \subseteq \mathrm{Ker}(T)$, which establishes the first conclusion.

To prove the second conclusion, we merely apply the first result to T^* instead of T, so that

$$\mathrm{Ker}(T^*) = \mathrm{R}(T^{**})^\perp = \mathrm{R}(T)^\perp.$$

Now, the third conclusion follows by taking the orthogonal complement of the second result. In particular, because $\mathrm{R}(T)$ is a closed subspace of $\mathcal{V}$, we find by Theorem 4.26

$$\mathrm{Ker}(T^*)^\perp = \mathrm{R}(T)^{\perp\perp} = \mathrm{R}(T).$$

The final conclusion follows by applying the third result to T^* rather than T. □

Comment. Though we assume $\mathrm{R}(T)$ is closed to obtain the last two conclusions, this is not necessary to arrive at a similar statement. If we remove this assumption, then they merely become

1. $\overline{\mathrm{R}(T)} = \mathrm{Ker}(T^*)^\perp$
2. $\overline{\mathrm{R}(T^*)} = \mathrm{Ker}(T)^\perp$

Comment. Notice that from the Decomposition Theorem (Theorem 4.25), the fundamental theorem now provides an immediate decomposition of $\mathcal{V}$ and $\mathcal{W}$ in terms of the kernel and range of a given operator T and its adjoint T^*, namely

$$\mathcal{V} = \mathrm{R}(T^*) \bigoplus \mathrm{R}(T^*)^\perp = \mathrm{Ker}(T) \bigoplus \mathrm{R}(T^*)$$

and

$$\mathcal{W} = \mathrm{R}(T) \bigoplus \mathrm{R}(T)^\perp = \mathrm{R}(T) \bigoplus \mathrm{Ker}(T^*).$$

These four subspaces of $\mathcal{V}$ and $\mathcal{W}$ are often referred to as the Four Fundamental subspaces induced by T.

We will now discuss two other direct consequences of the fundamental theorem - the Fredholm Alternative and the solution of Least Squares problems.

Theorem 5.10 (Fredholm Alternative)**.** Let $\mathcal{V}$ and $\mathcal{W}$ be Hilbert spaces and $T : \mathcal{V} \to \mathcal{W}$ be a bounded linear operator with $\mathrm{R}(T)$ closed. Given $w \in \mathcal{W}$, exactly one of the following must hold:

1. $T(v) = w$ has a solution $v \in \mathcal{V}$
2. $T^*(u) = 0$ has a non-trivial solution $u \in \mathcal{W}$ satisfying $\langle u, w \rangle \neq 0$.

Proof. Notice that the third conclusion of Theorem 5.9, namely

$$\mathrm{R}(T) = \mathrm{Ker}(T^*)^\perp$$

implies that $w \in \mathrm{R}(T)$ if and only if $w \in \mathrm{Ker}(T^*)^\perp$. Of course, in part (a) of the Fredholm Alternative $w \in \mathrm{R}(T)$, and thus $w \perp \mathrm{Ker}(T^*)$. Now, if $w \notin \mathrm{R}(T)$, then there exists an element $u \in \mathrm{Ker}(T^*)$ such that $w \not\perp u$ (and hence $u \neq 0$), which is exactly the statement of part (b). Thus, the Fredholm Alternative for finite dimensional vector spaces is merely a consequence of this theorem put into action, as any $w \in \mathcal{W}$ must satisfy either $w \in \mathrm{R}(T)$ or $w \notin \mathrm{Ker}(T^*)^\perp$. □

Example 41. Though we won't discuss it in detail, for any $\Omega \subset \mathbb{R}^3$ there is an important subspace of $L^2(\Omega)$ that we will denote by $\mathcal{H}$ for which the linear operator $T : \mathcal{H} \to L^2(\Omega)$ defined by

$$T[u] = \Delta u := \sum_{i=1}^{n} \partial_{x_i x_i} u$$

is bounded and $\mathrm{R}(T)$ is closed. Additionally, recall that by integration by parts T is self-adjoint so that $T^*[u] = T[u] = \Delta u$. Therefore, for any $f \in L^2(\Omega)$ we may apply Theorem 5.10 to find that one and only one of the following can be true:

1. $\Delta u = f$ has a solution $u \in \mathcal{H}$
2. $\Delta u = 0$ has non-trivial solution $u \in \mathcal{H}$ satisfying $\langle u, f \rangle \neq 0$

As you might imagine, this is a useful theorem that is often used for obtaining information about solutions of Laplace's or Poisson's equation.

Theorem 5.11 (Fredholm Alternative for Adjoint)**.** Let $\mathcal{V}$ and $\mathcal{W}$ be Hilbert spaces and $T : \mathcal{V} \to \mathcal{W}$ be a bounded linear operator. Given $w \in \mathcal{W}$, exactly one of the following must hold:

1. $T^*(w) = 0$
2. There is a non-trivial $z \in \mathcal{W}$ satisfying $\langle w, z \rangle \neq 0$ such that $T(v) = z$ has a solution $v \in \mathcal{V}$.

Proof. Notice that the second conclusion of Theorem 5.9, namely

$$\mathrm{Ker}(T^*) = R(T)^\perp$$

implies that $w \in \mathrm{Ker}(T^*)$ if and only if $w \in \mathrm{R}(T)^\perp$. Of course, in part (a) of the Fredholm Alternative $w \in \mathrm{Ker}(T^*)$, and thus $w \perp \mathrm{R}(T)$. Now, if $w \notin \mathrm{Ker}(T^*)$, then there exists an element $z \in \mathrm{R}(T)$ such that $z \not\perp w$, which is exactly the statement of part (b). □

Often, we wish to solve a linear system of (algebraic, integral, differential, or partial differential) equations, but find that the system is inconsistent, and thus possesses no solution. Hence, a natural question is - can we generalize the notion of solution to find at least one vector that satisfies a condition similar to solving the linear system. This condition will be exactly that of a least squares solution. As we will see later in the discussion, the QR Factorization can also be used to simplify such problems on finite-dimensional spaces, which arise naturally within a variety of disciplines, including but not limited to, statistics, image processing, and computational solutions of partial differential equations.

Definition 5.7. Let $\mathcal{V}$ and $\mathcal{W}$ be Hilbert spaces with $T : \mathcal{V} \to \mathcal{W}$ a bounded linear operator, and let $w \in \mathcal{W}$ be given. A vector $u \in \mathcal{V}$ is called a **least squares solution** of $T(v) = w$ if u satisfies

$$\|T(u) - w\|_{\mathcal{W}} \leq \|T(v) - w\|_{\mathcal{W}}$$

for every $v \in \mathcal{V}$.

If $T(v) = w$ is **consistent** (i.e., possesses a solution), then finding all least squares solutions is equivalent to finding all solutions of $T(v) = w$.

If $T(v) = w$ is **inconsistent** (i.e., has no solution), then $\|T(u) - w\|_{\mathcal{W}} > 0$.

From this definition, a least squares solution u minimizes the distance between $T(v)$ and w among all possible vectors $v \in \mathcal{V}$. Thus, if we can't find $v \in \mathcal{V}$ such that $T(v) = w$, which of course makes $T(v) - w = 0$, then we can do the next best thing, which is minimize the value of this difference.

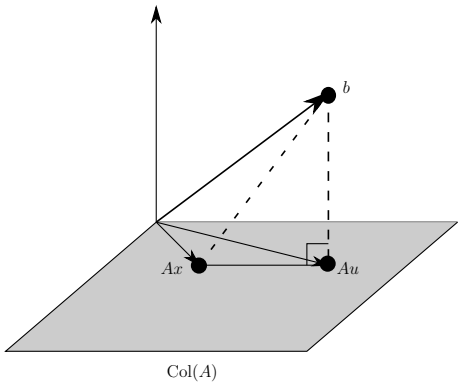


Figure 5.1: A representation of the projection of b onto $\mathrm{Col}(A)$, denoted by Au. Notice that $b - Au \perp \mathrm{Col}(A)$ and the distance $\|Au - b\|_2$ is minimal amongst all other vectors in $\mathrm{Col}(A)$, e.g. $\|Ax - b\|_2$.

Comment. For most problems of this type, we consider the finite-dimensional case $Ax = b$, in which $\mathcal{V} = \mathbb{R}^q$ and $\mathcal{W} = \mathbb{R}^p$ are both endowed with the standard inner product and $T(x) = Ax$ for some given $A \in \mathbb{R}^{p\times q}$ and $b \in \mathbb{R}^p$. In this case, the geometry is a bit easier to visualize (see Figure 5.1) using the projection onto $\mathrm{R}(T) = \mathrm{Col}(A)$. Due to the structure of the 2-norm on $\mathbb{R}^p$, the least squares definition is equivalent to

$$\|Au - b\|_2^2 \leq \sum_{i=1}^{p} \left[\left(\sum_{j=1}^{q} a_{ij} x_j \right) - b_i \right]^2$$

in this case, which demonstrates that such a solution minimizes a sum of squares, and hence motivates the name "least squares" solution.

Of course, finding a minimizer of a problem like this appears to be much more difficult than merely solving a linear system of equations. Fortunately, we can show that it is actually just as easy.

Theorem 5.12 (Normal Equations)**.** Let $w \in \mathcal{W}$ be given and assume $\mathrm{R}(T)$ is closed. Then, there exists a least squares solution $u \in \mathcal{V}$. Additionally, any $u \in \mathcal{V}$ is a least squares solution of $T(v) = w$ if and only if

$$(T^*T)(u) = T^*(w). \tag{5.1}$$

Proof. Let $w \in \mathcal{W}$ be given. Then, by the Projection Theorem (Theorem 4.24) with $M = \mathrm{R}(T)$ there is a unique $w^* \in \mathrm{R}(T)$ such that

$$\|w^* - w\|_{\mathcal{W}} = \inf_{z\in \mathrm{R}(T)} \|z - w\|_{\mathcal{W}} \tag{5.2}$$

and $w^* - w \in \mathrm{R}(T)^\perp$. Because $w^* \in \mathrm{R}(T)$, there exists a (not necessarily unique!) $u \in \mathcal{V}$ such that $T(u) = w^*$, and this u must be a least squares solution. Indeed, since $z \in \mathrm{R}(T)$ implies $T(v) = z$ for some $v \in \mathcal{V}$ and using (5.2), we find

$$\|T(u) - w\|_{\mathcal{W}} = \inf_{z\in \mathrm{R}(T)} \|z - w\|_{\mathcal{W}} = \inf_{v\in \mathcal{V}} \|T(v) - w\|_{\mathcal{W}} \leq \|T(v) - w\|_{\mathcal{W}}$$

for all $v \in \mathcal{V}$.

Now, by the Fundamental Theorem (Theorem 5.9), we know $\mathrm{R}(T)^\perp = \mathrm{Ker}(T)^*$. Thus, $w - w^* \in \mathrm{Ker}(T^*)$. Therefore, we have $w - T(u) \in \mathrm{Ker}(T^*)$ which means

$$T^*(w - T(u)) = 0$$

and by linearity of T^*, this becomes $T^*(T(u)) = T^*(w)$ which is exactly (5.1). Therefore, if $u \in \mathcal{V}$ is a least squares solution, it must satisfy (5.1).

To prove the reverse direction, if $u \in \mathcal{V}$ satisfies (5.1), then notice that $T(u) - w \in \mathrm{R}(T)^\perp$ from above. Then, given any $v \in \mathcal{V}$ we decompose the difference between $T(v)$ and w into two components, x and y, defined by

$$T(v) - w = \underbrace{T(v-u)}_{=:x} + \underbrace{T(u) - w}_{=:y}.$$

Notice that

$$\langle x, y \rangle = \langle T(v-u), T(u) - w \rangle = 0$$

because $T(u) - w \in \mathrm{R}(T)^\perp$ and $T(v-u) \in \mathrm{R}(T)$. With this, we find by the Pythagorean Theorem (cf. Problem 4.16)

$$\begin{aligned} \|T(v) - w\|_{\mathcal{W}}^2 &= \langle T(v) - w, T(v) - w \rangle_{\mathcal{W}} \\ &= \langle x + y, x + y \rangle_{\mathcal{W}} \\ &= \|x\|_{\mathcal{W}}^2 + \|y\|_{\mathcal{W}}^2 \\ &\geq \|y\|_{\mathcal{W}}^2 \\ &= \|T(u) - w\|_{\mathcal{W}}^2. \end{aligned}$$

Since v was arbitrary, this inequality holds for all $v \in \mathcal{V}$, and thus u is a least squares solution of $T(v) = w$.

□

Comment. Because the adjoint operator is exactly the generalization of the transpose, the condition (5.1) in the finite dimensional case is exactly

$$A^T A u = A^T b,$$

which you may have seen in a Linear Algebra course.

Comment. We can define a generalized notion of least squares solution in Banach spaces or using other norms (e.g., $\|Ax - b\|_1$ or $\|Ax - b\|_\infty$ in the finite-dimensional case), but a result analogous to Theorem 5.12 may not exist, and this makes the associated minimization problem much more difficult to solve.

Comment. A well-known class of computational methods for solving PDEs, known as Least Squares Finite Element Methods (LSFEMs), is based exactly on the formulation of a least squares problem in a Hilbert space. In this case, $\mathcal{V}$ is often a particular subspace of $L^2(a, b)$ and the PDEs are solved by minimizing the norm of a certain operator associated to the given PDE.

In the case that infinitely many solutions of this problem arise, we can also determine the structure of the associated degrees of freedom without much work. In particular, as the next result displays, additional solutions are constructed by moving through $\text{Ker}(T)$.

Theorem 5.13. If $u \in \mathcal{V}$ is a least squares solution of $T(v) = w$ and $z \in \text{Ker}(T)$, then $u + \alpha z$ is a least squares solution of $T(v) = w$ for any $\alpha \in \mathbb{K}$.

Proof. Let $\alpha \in \mathbb{K}$ be given. Since $z \in \text{Ker}(T)$, we see that $T(z) = 0$. Thus, we compute

$$T(u + \alpha z) = T(u) + \alpha T(z) = T(u).$$

Therefore, we find

$$T^*(T(u + \alpha z)) = T^*(T(u)) = T^*(w)$$

and $u + \alpha z$ satisfies the normal equations. Therefore, by Theorem 5.12, $u + \alpha z$ is a least squares solution of $T(u) = w$. □

As new least squares solutions are generated by $\text{Ker}(T)$, we can clearly deduce the following result.

Corollary 5.1. Let $T : \mathcal{V} \to \mathcal{W}$ be a bounded linear operator and $w \in \mathcal{W}$ be given. Then,

1. If $\text{Ker}(T) = \{0\}$, then $T(v) = w$ has exactly one least squares solution, while
2. If $\text{Ker}(T) \neq \{0\}$, then $T(v) = w$ has infinitely many least squares solutions.

Turning to the finite-dimensional problem and in view of this corollary, the corresponding condition $\text{Nul}(A) = \{0\}$ can be challenging to directly verify. Still we can find specific (and easy to check) conditions on the matrix A that indicate how many least squares solutions will exist.

Theorem 5.14. Let $A \in \mathbb{R}^{p \times q}$ and $b \in \mathbb{R}^p$ be given and consider solving the least squares problem $Ax = b$. Then,

1. If $\text{rank}(A) = q$ then $Ax = b$ has exactly one least squares solution, while
2. If $\text{rank}(A) < q$ then $Ax = b$ has infinitely many least squares solutions.

Proof. This result follows immediately from Corollary 5.1 and the Rank-Nullity Theorem, but we include a direct proof for completeness.

From Theorem 2.7, $\text{rank}(A) = q$ if and only if $A^T A$ is nonsingular, and by the IMT, this is true if and only if $A^T Ax = A^T b$ has a unique solution. Analogously, we may apply this same argument in the opposite direction, so that Theorem 2.7, tell us that $\text{rank}(A) < q$ if and only if $A^T A$ is singular, and by the IMT, this is true if and only if $A^T Ax = A^T b$ has infinitely many solutions. □

Though we have simplified the minimization problem to one with which we have greater familiarity, namely the solution of a linear system, it should be noted that this simplification is not without its own difficulties. For instance, if A is a large matrix, then computing and storing $A^T A$ may be expensive, and computing the solution of the normal equations even more so. Fortunately, we can simplify this problem even further using the QR Factorization that we previously derived in Theorem 4.22.

Theorem 5.15. A vector $u \in \mathbb{R}^q$ is a least squares solution of $Ax = b$ if and only if

$$Ru = Q^T b$$

where Q and R comprise a QR Factorization of A; namely, they satisfy $A = QR$ and the properties stated in Theorem 4.22.

Proof. Evoking Theorem 5.12, we merely need to transform the normal equations to prove the theorem. By Theorem 4.22, we can write $A = QR$ where $Q \in \mathbb{R}^{p\times k}$ has orthonormal columns and $R \in \mathbb{R}^{k\times q}$ satisfies $\text{rank}(R) = k$. Because Q has orthonormal columns, we see that

$$q_i^T q_j = \begin{cases} 1, & \text{if } i = j \\ 0, & \text{if } i \neq j \end{cases}$$

for all $i, j = 1, ..., k$ and thus $Q^T Q = \mathbb{I}$. Then, we find

$$\begin{aligned} A^T Au = A^T b &\iff (QR)^T (QR)u = (QR)^T b \\ &\iff R^T Q^T QRu = R^T Q^T b \\ &\iff R^T Ru = R^T Q^T b \end{aligned}$$

Now, by Problem 4.17, since R has full row rank (that is, $R \in \mathbb{R}^{k\times q}$ with $\text{rank}(R) = k$), there is $X \in \mathbb{R}^{q\times k}$ such that $RX = \mathbb{I}$. Taking the transpose of this equation yields $X^T R^T = \mathbb{I}^T = \mathbb{I}$. Hence, we multiply the above vector equation by X^T on the left and use this property to find

$$Ru = Q^T b.$$

Of course, it's important to verify that this linear system does indeed possess a solution. Certainly, we let $c = Q^T b$ and attempt to determine whether or not $Ru = c$ possesses a solution. Of course, $\text{rank}([R|c]) \leq k$ because this augmented matrix possesses only k rows (as $R \in \mathbb{R}^{k\times q}$). Additionally, due to the augmented structure of the matrix we see that $\text{rank}([R|c]) \geq \text{rank}(R) = k$. Combining these inequalities, we find

$$\text{rank}(R) = \text{rank}([R|c]) = k,$$

and by the Rank-Solvability Theorem (Theorem 2.5), the system is consistent. Finally, if $k = q$, we know that the solution is, in fact, unique, while if $k < q$ there will be infinitely many least squares solutions. □

Example 42. Define the 3×3 matrix A and vector b by

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 0 \\ 1 & 2 & 0 \end{bmatrix}, \qquad b = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix},$$

and find all least squares solutions of $Ax = b$ using Theorem 5.15.

From Example 31, we can decompose A into

$$A = \underbrace{\begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \end{bmatrix}}_{Q} \underbrace{\begin{bmatrix} \sqrt{3} & 2\sqrt{3} & \sqrt{3} \\ 0 & 0 & \sqrt{6} \end{bmatrix}}_{R}.$$

Next, we compute

$$c = Q^T b = \begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ \frac{2}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{2}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} \end{bmatrix}$$

and solve $Ru = c$ or

$$\begin{bmatrix} \sqrt{3} & 2\sqrt{3} & \sqrt{3} \\ 0 & 0 & \sqrt{6} \end{bmatrix} u = \begin{bmatrix} \frac{2}{\sqrt{3}} \\ \frac{1}{\sqrt{6}} \end{bmatrix},$$

which, after some algebra yields

$$\begin{cases} u_1 + 2u_2 = \dfrac{1}{2} \\ \qquad\quad u_3 = \dfrac{1}{6} \end{cases}$$

or, the parametrized family of solutions

$$u = s \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} + \begin{bmatrix} \frac{1}{2} \\ 0 \\ \frac{1}{6} \end{bmatrix}$$

are all least squares solutions for any $s \in \mathbb{R}$. Obviously, we find infinitely many solutions because $q = 3$ while $k = 2$ so that $k < q$. Furthermore, we see that these solutions are constructed by moving through $\text{Nul}(A)$ as

$$A \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

and thus

$$\begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} \in \text{Nul}(A).$$

In the next chapter, we will explore some specific applications of least squares problems. In particular, we will discuss the use of Linear Regression for constructing linear statistical models for prediction and ranking systems that are generally used for sports and other competitions.

5.4 Norms of Linear Operators

In this section, we let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces (not necessarily possessing an inner product).

Definition 5.8. Assume $T : \mathcal{V} \to \mathcal{W}$ is linear and bounded. Then, the **norm of** T is defined by

$$\|T\| = \sup_{v \in \mathcal{V} \setminus \{0\}} \frac{\|T(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}}$$

In some contexts, it is possible to define other norms of T, and thus this particular norm is often referred to as the **operator norm of** T.

Comment. A few remarks concerning the norm of a linear operator:

1. If $\dim(\mathcal{V}) < \infty$ and $\dim(\mathcal{W}) < \infty$, the operator norm is defined analogously, but with the supremum replaced by the maximum.

2. This definition is equivalent to defining

$$\|T\| = \sup_{\|v\|_{\mathcal{V}}=1} \|T(v)\|_{\mathcal{W}}.$$

Indeed, for any $v \in \mathcal{V} \setminus \{0\}$ if we let $u = \frac{v}{\|v\|_{\mathcal{V}}}$, then

$$\|T(u)\|_{\mathcal{W}} = \left\| T\left(\frac{v}{\|v\|_{\mathcal{V}}}\right) \right\|_{\mathcal{W}} = \left\| \frac{1}{\|v\|_{\mathcal{V}}} T(v) \right\|_{\mathcal{W}} = \frac{\|T(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}}.$$

Thus, maximizing the right side over all such v is equivalent to maximizing the left side over all such u, which must be a unit vector.

Example 43. Recall that $\ell^2(\mathbb{R})$ is a Hilbert space with norm

$$\|x\|_2^2 = \sum_{k=1}^{\infty} x_k^2$$

and consider the left shift operator $L : \ell^2 \to \ell^2$ defined by

$$L(x) = (x_2, x_3, ...)$$

for every $x = (x_1, x_2, x_3, ...) \in \ell^2$. Then, notice that for any $x \in \ell^2$

$$\|L(x)\|_2^2 = \sum_{k=2}^{\infty} x_k^2 \leq \sum_{k=1}^{\infty} x_k^2 = \|x\|_2^2$$

so that

$$\frac{\|L(x)\|_2}{\|x\|_2} \leq 1.$$

Since this holds for all $x \in \ell^2$, we find $\|L\| \leq 1$. Additionally, if we consider $e_3 \in \ell^2$ defined by

$$e_3 = (0, 0, 1, 0, 0, ...)$$

then $\|e_3\|_2 = 1$, $\|L(e_3)\|_2 = 1$, and thus

$$\frac{\|L(e_3)\|_2}{\|e_3\|_2} = 1.$$

Therefore, the supremum satisfies

$$\|T\| = \sup_{\|v\|_2=1} \|T(v)\|_2 \geq 1,$$

and combining this with the previous inequality, we conclude $\|T\| = 1$.

Next, we show that the operator norm is, in fact, a norm defined on the space of linear operators from $\mathcal{V}$ to $\mathcal{W}$.

Theorem 5.16. Assume $T, T_1, T_2 : \mathcal{V} \to \mathcal{W}$ are linear and bounded. Then,

(a) $\|0\| = 0$, and if $T \neq 0$ then $\|T\| > 0$.

(b) For all $\alpha \in \mathbb{K}$, we have $\|\alpha T\| = |\alpha|\|T\|$

(c) The operator norm satisfies the triangle inequality, namely

$$\|T_1 + T_2\| \leq \|T_1\| + \|T_2\|.$$

Thus, $\|T\|$ is a norm.

(d) For every $v \in \mathcal{V}$, we have

$$\|T(v)\|_{\mathcal{W}} \leq \|T\|\|v\|_{\mathcal{V}}.$$

Proof. The first two conclusions are straightforward, so we focus on proving only the last two conclusions. To prove part (c), we first write

$$\|T_1 + T_2\| = \sup_{v \in \mathcal{V}\setminus\{0\}} \frac{\|T_1(v) + T_2(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}}$$

Thus, using the triangle inequality for the norm $\|\cdot\|_{\mathcal{W}}$ we find

$$\begin{aligned}\|T_1 + T_2\| &\leq \sup_{v \in \mathcal{V}\setminus\{0\}} \left(\frac{\|T_1(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} + \frac{\|T_2(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} \right) \\ &\leq \sup_{v \in \mathcal{V}\setminus\{0\}} \frac{\|T_1(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} + \sup_{v \in \mathcal{V}\setminus\{0\}} \frac{\|T_2(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} \\ &= \|T_1\| + \|T_2\|.\end{aligned}$$

To prove the final conclusion, we let $u \in \mathcal{V}$ with $u \neq 0$ be given and write

$$\frac{\|T(u)\|_{\mathcal{W}}}{\|u\|_{\mathcal{V}}} \leq \sup_{v \in \mathcal{V}\setminus\{0\}} \frac{\|T(v)\|_{\mathcal{W}}}{\|v\|_{\mathcal{V}}} = \|T\|.$$

Multiplying by $\|u\|_{\mathcal{V}}$ then provides the result for any $u \in \mathcal{V}$. □

When considering $\dim(\mathcal{V}) < \infty$ and $\dim(\mathcal{W}) < \infty$, we can return to our prototypical example of a linear operator, namely $T : \mathbb{R}^q \to \mathbb{R}^p$ defined by $T(v) = Av$ for some $A \in \mathbb{R}^{p\times q}$. Of course, since matrices are just special varieties of linear operators, this definition remains valid for them as well.

Definition 5.9. Assume $A \in \mathbb{R}^{p\times q}$, then the **operator norm of** A is defined by

$$\|A\| = \max_{x \in \mathbb{R}^q\setminus\{0\}} \frac{\|Ax\|_{\mathbb{R}^p}}{\|x\|_{\mathbb{R}^q}}.$$

A minor notational difficulty arises here, in that there are infinitely many norms on $\mathbb{R}^n$ for fixed $n \in \mathbb{N}$; for instance, we saw earlier that

$$\|x\|_p = \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p}$$

is a norm on $\mathbb{R}^n$ for any choice of $p \in \mathbb{N}$, as is the maximum norm $\|\cdot\|_\infty$. For this reason, when dealing with matrix norms we will use subscripts to denote the specific p-norm and not the the vector space or its dimension - for example,

$$\|A\|_2 = \max_{x \in \mathbb{R}^q \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2}$$

and

$$\|A\|_\infty = \max_{x \in \mathbb{R}^q \setminus \{0\}} \frac{\|Ax\|_\infty}{\|x\|_\infty}.$$

Of course, the situation could even be more troublesome than we've described, as the norm on the codomain $\mathbb{R}^p$ does not need to be the same as the norm on the domain $\mathbb{R}^q$, but we will not encounter this situation in the future.

For two specific matrix norms, $\|A\|_1$ and $\|A\|_\infty$, we can actually construct a formula for these quantities only in terms of entries of the given matrix A.

Theorem 5.17. Let $A \in \mathbb{R}^{p \times q}$ be given with entries a_{ij}. Then, we have

1. $\|A\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^{q} |a_{ij}|$

2. $\|A\|_1 = \max_{1 \leq j \leq q} \sum_{i=1}^{p} |a_{ij}|$

Proof. We will prove only the first conclusion, as the second portion is left as a homework exercise (cf. Problem 5.5). Define the number

$$\alpha = \max_{1 \leq i \leq p} \sum_{j=1}^{q} |a_{ij}|.$$

In order to prove $\|A\|_\infty = \alpha$, the strategy will be to show both $\|A\|_\infty \leq \alpha$ and $\|A\|_\infty \geq \alpha$. This will be done by first establishing the inequality

$$\|Ax\|_\infty \leq \alpha \|x\|_\infty$$

for every $x \in \mathbb{R}^q \setminus \{0\}$, which yields $\|A\|_\infty \leq \alpha$, and then later proving

$$\|Ax\|_\infty \geq \alpha \|x\|_\infty$$

for some choice of $x \in \mathbb{R}^q \setminus \{0\}$, which will guarantee $\|A\|_\infty \geq \alpha$.

We focus on establishing the upper bound on $\|A\|_\infty$ first. In particular, we let $x \in \mathbb{R}^q \setminus \{0\}$ be given and write

$$\|Ax\|_\infty = \max_{1 \leq i \leq p} \left| \sum_{j=1}^{q} a_{ij} x_j \right| = \left| \sum_{j=1}^{q} a_{kj} x_j \right|$$

for some $k = 1, \ldots, p$ because the maximum over a finite set must be attained. Using the Triangle Inequality, this becomes

$$\begin{aligned}
\|Ax\|_\infty &\leq \sum_{j=1}^{q} |a_{kj}||x_j| \\
&\leq \sum_{j=1}^{q} |a_{kj}| \left(\max_{1\leq j\leq q} |x_j| \right) \\
&= \left(\sum_{j=1}^{q} |a_{kj}| \right) \|x\|_\infty \\
&\leq \left(\max_{1\leq i\leq p} \sum_{j=1}^{q} |a_{ij}| \right) \|x\|_\infty \\
&= \alpha \|x\|_\infty,
\end{aligned}$$

and thus dividing by $\|x\|_\infty \neq 0$ we have

$$\frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \alpha$$

for every $x \in \mathbb{R}^q \setminus \{0\}$. Since α is a uniform (i.e., independent of the choice of x) upper bound, we can take the maximum of both sides over all $x \in \mathbb{R}^q \setminus \{0\}$ and find

$$\|A\|_\infty = \max_{x\in\mathbb{R}^q\setminus\{0\}} \frac{\|Ax\|_\infty}{\|x\|_\infty} \leq \alpha.$$

Next, we show the lower bound on $\|A\|_\infty$. As before, because the maximum is taken over a finite set, we know that $\alpha = \max_{1\leq i\leq p} \sum_{j=1}^{q} |a_{ij}|$ must attain this maximum value at some $k = 1, \ldots, p$ so that we can write

$$\alpha = \sum_{j=1}^{q} |a_{kj}|.$$

Given this value of k, we define $y \in \mathbb{R}^q \setminus \{0\}$ by $y = \begin{bmatrix} y_1 \\ \vdots \\ y_q \end{bmatrix}$ where each entry satisfies

$$y_j = \begin{cases} \dfrac{|a_{kj}|}{a_{kj}}, & \text{if } a_{kj} \neq 0 \\ 1, & \text{else.} \end{cases}$$

From this definition, we see that

$$a_{kj}y_j = \begin{cases} |a_{kj}|, & \text{if } a_{kj} \neq 0 \\ 0, & \text{else.} \end{cases}$$

which can be expressed merely as $a_{kj}y_j = |a_{kj}|$. Additionally, it follows from the definition of y that $|y_j| = 1$ for all $j = 1, ..., q$ and thus $\|y\|_\infty = 1$. Now, notice that

$$\|Ax\|_\infty = \max_{1\leq i\leq p} \left| \sum_{j=1}^{q} a_{ij}x_j \right| \geq \left| \sum_{j=1}^{q} a_{ij}x_j \right|$$

for any $1 \leq i \leq p$ and $x \in \mathbb{R}^q \setminus \{0\}$. Therefore, we have

$$\|Ay\|_\infty \geq \left| \sum_{j=1}^{q} a_{kj} y_j \right| = \left| \sum_{j=1}^{q} |a_{kj}| \right| = \sum_{j=1}^{q} |a_{kj}| = \alpha = \alpha \|y\|_\infty.$$

Hence, $\|Ay\|_\infty \geq \alpha \|y\|_\infty$ for some $y \in \mathbb{R}^q \setminus \{0\}$. As before, we can divide by $\|y\|_\infty \neq 0$ and note that

$$\|A\|_\infty = \max_{x \in \mathbb{R}^q \setminus \{0\}} \frac{\|Ax\|_\infty}{\|x\|_\infty} \geq \frac{\|Ay\|_\infty}{\|y\|_\infty} \geq \alpha.$$

With the lower bound on $\|A\|_\infty$, we now have $\|A\|_\infty = \alpha$ and the proof is complete. □

Example 44. Consider the 2×2 matrix $A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$. Using Theorem 5.17, we can easily compute these norms of A, so that

$$\|A\|_\infty = \max_{1 \leq i \leq p} \sum_{j=1}^{q} |a_{ij}| = \max_{1 \leq i \leq p} (\underbrace{|a_{11}| + |a_{12}|}_{i=1}, \underbrace{|a_{21}| + |a_{22}|}_{i=2}) = \max\{3, 7\} = 7$$

and

$$\|A\|_1 = \max_{1 \leq j \leq q} \sum_{i=1}^{p} |a_{ij}| = \max_{1 \leq j \leq q} (\underbrace{|a_{11}| + |a_{21}|}_{j=1}, \underbrace{|a_{12}| + |a_{22}|}_{j=2} = \max\{4, 6\} = 6.$$

In the next chapter we will investigate some applications of least squares problems to linear regression and ranking systems. After that, we discuss properties of special types of linear operators on Hilbert spaces, including their eigenvalues and eigenvectors. This will have specific implications regarding how these operators might be represented or decomposed in terms of their associated eigenvectors.

Exercises - Linear Operators

Problem 5.1. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces with $\dim(\mathcal{V}) < \infty$, and let $T : \mathcal{V} \to \mathcal{W}$ be a linear operator. Assume that $\{w_1, ..., w_k\}$ is a basis for $\mathrm{R}(T)$ and let $B_1 = \{v_1, .., v_k\}$ where $T(v_j) = w_j$ for all $j = 1, ..., k$. Prove that if $B_2 = \{u_1, ..., u_n\}$ is a basis for $\mathrm{Ker}(T)$, then $B := B_1 \cup B_2$ is basis for $\mathcal{V}$.

Problem 5.2. Let $p, q \in \mathbb{N}$ and $A \in \mathbb{C}^{p\times q}$ be given. Prove that

$$\mathrm{rank}(A^H A) = \mathrm{rank}(A).$$

Hint: First show that $\mathrm{Ker}(A^H A) = \mathrm{Ker}(A)$.

Problem 5.3. Let $\mathcal{V}$, $\mathcal{W}$, and $\mathcal{Z}$ be vector spaces over the same field $\mathbb{K}$ with linear operators $S : \mathcal{W} \to \mathcal{Z}$ and $T : \mathcal{V} \to \mathcal{W}$. Define the composition $ST : \mathcal{V} \to \mathcal{Z}$ for every $v \in \mathcal{V}$ by

$$(ST)(v) = S(T(v)).$$

(a) Show that ST is linear.

(b) Assume that S and T are onto and $\mathrm{Ker}(S) = \mathrm{Ker}(T) = \{0\}$ so that both S and T have an inverse. Show that ST has an inverse and for every $z \in \mathcal{Z}$

$$(ST)^{-1}(z) = (T^{-1}S^{-1})(z).$$

Problem 5.4. Let $\mathcal{V}$ and $\mathcal{W}$ be inner product spaces over the same field $\mathbb{K}$. For any $M \subset \mathcal{V}$, we define

$$M^{\perp} = \{v \in \mathcal{V} : v \perp w \text{ for every } w \in M\}.$$

Assume that $T : \mathcal{V} \to \mathcal{W}$ is a bounded linear operator.

(a) Show directly that $R(T^*) \subset \mathrm{Ker}(T)^{\perp}$ and then use this to show that $R(T) \subset \mathrm{Ker}(T^*)^{\perp}$.

(b) Let $p, q \in \mathbb{N}$ and $A \in \mathbb{R}^{p\times q}$ be given. Formulate the statements from (a) in terms of the operator $T : \mathbb{R}^q \to \mathbb{R}^p$ defined by $T(x) = Ax$.

Problem 5.5. Let $p, q \in \mathbb{N}$ and recall for $A \in \mathbb{R}^{p\times q}$, the norm $\|A\|_1$ is defined by

$$\|A\|_1 = \max_{x\in\mathbb{R}^q\setminus\{0\}} \frac{\|Ax\|_1}{\|x\|_1}.$$

Show that $\|A\|_1 = \max\limits_{1\le j\le q} \sum\limits_{i=1}^{p} |a_{ij}|$.

Problem 5.6. Let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces with $\dim(\mathcal{V}) < \infty$ and assume $T : \mathcal{V} \to \mathcal{W}$ is a linear operator. Prove that T is bounded.
Hint: Lemma 4.11 might be useful.

Problem 5.7. Let $\mathcal{V}$ and $\mathcal{W}$ be normed spaces and assume $T : \mathcal{V} \to \mathcal{W}$ is a linear operator. We say that T is **continuous** on $\mathcal{V}$ if for every $v_0 \in \mathcal{V}$ and $\epsilon > 0$, there exists $\delta > 0$ such that

$$\|v - v_0\|_{\mathcal{V}} < \delta \quad \text{implies} \quad \|T(v) - T(v_0)\|_{\mathcal{W}} < \epsilon.$$

Prove that T is continuous if and only if T is bounded.

Problem 5.8. Let $\mathcal{V}, \mathcal{W}, \mathcal{Z}$ be Hilbert spaces and assume $T_1 : \mathcal{V} \to \mathcal{W}$ and $T_2 : \mathcal{W} \to \mathcal{Z}$ are bounded linear operators. Define $S : \mathcal{V} \to \mathcal{Z}$ by their composition $S = T_2 T_1$, namely

$$S(v) = T_2(T_1(v))$$

for all $v \in \mathcal{V}$.

(a) Show that S is bounded and satisfies

$$\|S\| \leq \|T_1\| \cdot \|T_2\|.$$

(b) Show that S has an adjoint S^* and

$$S^* = T_1^* T_2^*.$$

Problem 5.9. Let $\mathcal{V}$ be a Hilbert space and assume that $T : \mathcal{V} \to \mathcal{V}$ is a bounded linear operator. Show that

$$\|T^*\| = \|T\|.$$

Problem 5.10. Construct a normalized QR Factorization of A and use it to solve the least squares problem $Ax = b$, where

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 1 & -1 \end{bmatrix}, \qquad b = \begin{bmatrix} 2 \\ 3 \\ 0 \end{bmatrix}.$$

Chapter 6

Application: Linear Regression & Ranking

In this chapter, we will explore a few different applications of the least squares problems described in the last section. These include linear regression, i.e. fitting a linear statistical model to given data, and ranking algorithms.

6.1 Linear Regression

In general, solving least squares problems for linear statistical models allow us to

1. Spot qualitative trends in data
2. Construct models of the form

   $$y = x \cdot \beta + \varepsilon$$

   where $y, \varepsilon \in \mathbb{R}$ and $x, \beta \in \mathbb{R}^p$, which represents a simple linear regression
3. Construct models of the form

   $$y = X\beta + \varepsilon,$$

   where $y, \varepsilon \in \mathbb{R}^p$, $X \in \mathbb{R}^{p \times q}$, and $\beta \in \mathbb{R}^q$, which represents a multiple linear regression
4. Predict future outcomes from data, as output from the linear model.

Example 45. Consider the growth of the total population of the United States between 1950 and 2000. The following table provides data accumulated from each decade:

t (year)	y (population)	s
1950	150.697	0
1960	179.323	0.2
1970	203.212	0.4
1980	226.505	0.6
1990	249.633	0.8
2000	281.422	1

Here, the variable t represents the year, y the associated population (in millions), and we let

$$s = \frac{1}{50}(t - 1950)$$

to rescale time so that $0 \leq s \leq 1$. We assume a cubic interpolation model of the form

$$y = \beta_0 + \beta_1 s + \beta_2 s^2 + \beta_3 s^3 = \underbrace{\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{bmatrix}}_{\beta} \cdot \underbrace{\begin{bmatrix}1\\ s\\ s^2\\ s^3\end{bmatrix}}_{x}$$

which is a **linear model in** β expanded about the standard basis for $\mathbb{P}^3$. The data generates points in a linear system as the input and output values for s and y from the table give rise to the six equations:

$$\begin{aligned}
\beta_0 + \beta_1 \cdot 0 + \beta_2 \cdot 0^2 + \beta_3 \cdot 0^3 &= 150.697\\
\beta_0 + \beta_1 \cdot 0.2 + \beta_2 \cdot (0.2)^2 + \beta_3 \cdot (0.2)^3 &= 179.323\\
\vdots \qquad\qquad & \vdots \qquad\qquad \vdots\\
\vdots \qquad\qquad & \vdots \qquad\qquad \vdots\\
\beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot (1)^2 + \beta_3 \cdot (1)^3 &= 281.422
\end{aligned}$$

Of course, this yields six equations for four variables, and that does not bode well for a unique solution. Indeed, it turns out that these points do not lie on a unique cubic curve. Hence, we reformulate this as a least squares problem and solve that instead; in particular, we cannot find $\beta \in \mathbb{R}^4$ satisfying

$$A\beta = y$$

or

$$\underbrace{\begin{bmatrix}1 & 0 & 0 & 0\\ 1 & 0.2 & 0.2^2 & 0.2^3\\ 1 & 0.4 & 0.4^2 & 0.4^3\\ 1 & 0.6 & 0.6^2 & 0.6^3\\ 1 & 0.8 & 0.8^2 & 0.8^3\\ 1 & 1 & 1^2 & 1^3\end{bmatrix}}_{A} \underbrace{\begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\end{bmatrix}}_{\beta} = \underbrace{\begin{bmatrix}150.697\\ 179.323\\ 203.212\\ 226.505\\ 249.633\\ 281.422\end{bmatrix}}_{y}.$$

So, we instead compute the QR factorization of A in MATLAB to find

$$Q = \begin{bmatrix}
-0.4082 & -0.5976 & 0.5455 & -0.3727 & -0.1733 & -0.0983\\
-0.4082 & -0.3586 & -0.1091 & 0.5217 & 0.4954 & 0.4186\\
-0.4082 & -0.1195 & -0.4364 & 0.2981 & -0.2492 & -0.6911\\
-0.4082 & 0.1195 & -0.4364 & -0.2981 & -0.4925 & 0.5451\\
-0.4082 & 0.3586 & -0.1091 & -0.5217 & 0.6171 & -0.1995\\
-0.4082 & 0.5976 & 0.5455 & 0.3727 & -0.1976 & 0.0253
\end{bmatrix}$$

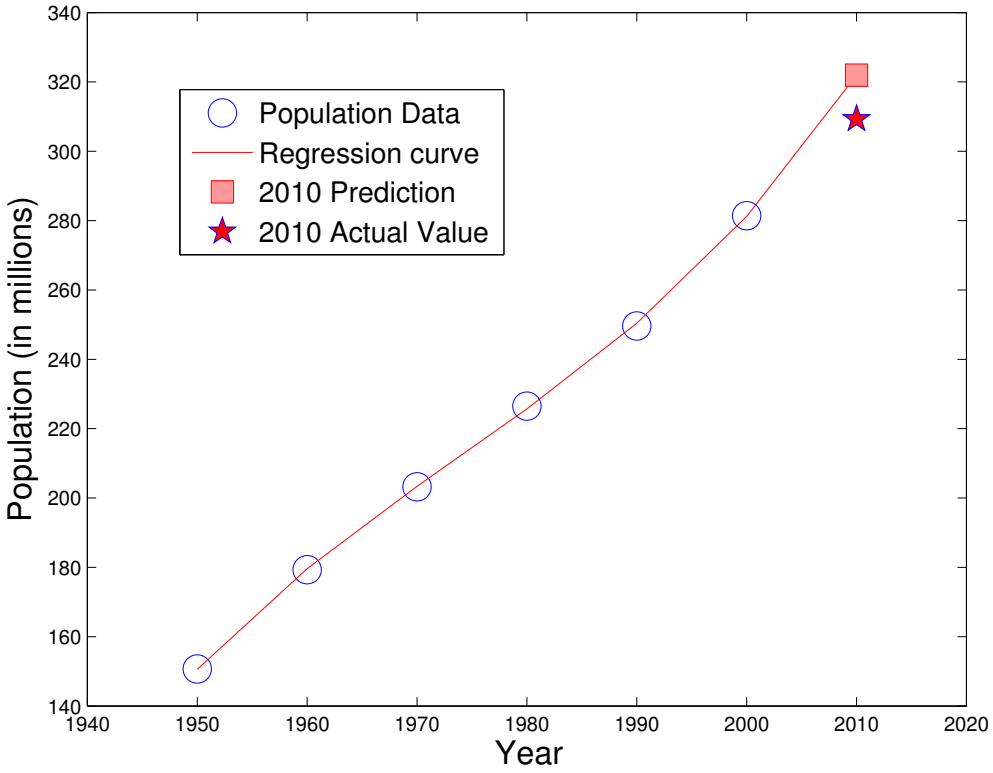


Figure 6.1: Plot of regression curve from U.S. Population example.

and

$$R = \begin{bmatrix} -2.4495 & -1.2247 & -0.8981 & -0.7348 \\ 0 & 0.8367 & 0.8367 & 0.7965 \\ 0 & 0 & 0.2444 & 0.3666 \\ 0 & 0 & 0 & 0.0644 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

then solve

$$R\beta = Q^T y,$$

which is also performed using MATLAB. The result is

$$\beta = \begin{bmatrix} 150.577 \\ 164.433 \\ -112.845 \\ 79.031 \end{bmatrix}$$

which yields the model

$$y(s) = 150.577 + 164.433s - 112.845s^2 + 79.031s^3$$

and the curve in Figure 6.1 shows the increasing trend in the population. As a side note, we mention that A has a special structure that arises in polynomial interpolation problems, and is called a **Vandermonde matrix**. Additionally, we can compute an estimate for the population in the year 2010 using

$$y(1.2) = 321.978.$$

The true population value was 309.3 million, which means the relative error in our approximation was

$$\frac{321.978 - 309.3}{309.3} \approx 4.1\%$$

and this is appears to have been a pretty good estimate with a mere six data points.

6.2 Ranking Systems

Given a number of items with particular value scores, we might wish to rank them. Examples of this type of situation include college rankings (e.g., U.S. News & World Report), movie suggestions (Netflix similarity scores), and sports rankings (e.g., BCS, FBS, Top 25, or chess). We will focus on the last of these for some examples of ranking systems that may incorporate least squares problems.

Let's consider ranking a collection of teams (say, football teams) who have played a certain number of games. In particular, we consider a few common ranking methods:

1. Winning Percentage

For each team $k = 1, ..., N$, we compute their associated winning percentage by dividing their wins by the total number of games:

$$W_k = \frac{w_k}{g_k}$$

where w_k represents the number of wins by team k and g_k represents the number of games played by team k.

This has a number of advantages, in that it's quite easy to compute and provides a numerical ordering (with ties, possibly) of the teams, but there are also disadvantages. For instance, W_k does not account for differences in schedule, and within a sport like college football this can be a large factor with so many teams and so few games. Additionally, this metric doesn't necessarily account for the number of games that have been played. In this way, an undefeated $20 - 0$ team and a $1 - 0$ team have identical winning percentages, though certainly the former team has displayed a greater ability to win games.

2. Predefined Point Structure

For each team $k = 1, ..., N$, we compute their number of points using a prescribed value associated to wins, losses, and perhaps other outcomes. Both the National Hockey League and English Premier League (soccer) utilize such a system, and we'll use the latter as an example of how the point total of a team is computed. Namely, each team receives three points for a victory, zero points for a loss, and one point for a draw:

$$P_k = 3w_k + t_k$$

where w_k and t_k represent the number of wins and draws by team k, respectively.

As an advantage over winning percentage, leagues which allow game outcomes other than wins and losses (e.g., ties or "overtime losses" in the NHL) can

associate a non-zero value to such an outcome. Of course, this ranking structure may incentivize draws, which has been one large critique over the years, and may disproportionately weight narrow victories relative to decisive ones, as the value associated to a win is the same regardless of the difference in score.

3. Overall Point Differential

For each team $k = 1, ..., N$, we compute their associated point differential by subtracting the overall number of points/goals surrendered (sometimes referred to as "points/goals against") during all games from the overall number of points/goals scored (sometimes referred to as "points/goals for") during all games:

$$D_k = \left(\sum_{j=1}^{g_k} GF_j\right) - \left(\sum_{j=1}^{g_k} GA_j\right)$$

where g_k represents the number of games played by team k, while GF_j and GA_j represent the number of goals for, respectively against, within game j.

Point differential is useful in that it provides a better understanding of a team's "strength of victory", so that narrow victories and defeats are not awarded or penalized as severely as routs. Contrastingly, ranking teams by point differential may incentivize "running up the score", in which a team that is soundly winning a game may display a lack of sportsmanship by embarrassing their opponents in an effort to increase the point differential. This was certainly an issue that arose in college football rankings when this metric was used an essential ingredient in the ranking formula. In addition, this metric is biased against teams who have played less games than others, as they would have had fewer opportunities to amass a large point differential.

4. Approximate Point Differential (Massey index)

Since the idea of ranking teams by point differential has a number of advantages, it may be possible to alter the original method to remove some of its shortcomings, and this is exactly the exercise undertaken by a mathematician named Ken Massey. In particular, we would like to construct a ranking method that is transitive in point differential, meaning that if Team A beats Team B by 10 points, and Team B beats Team C by 5 points, then we should expect that Team A would beat Team C by 15 points. Not only would this ranking then provide an idea of who should win when two teams play, but it would also predict the margin of victory (which is a particularly crucial number for many people who take an interest in college and professional sports - wink wink). Of course, if Team A and C do play and the result is not exactly as predicted by the ranking model, then the (linear) system of equations we would use to determine their rankings would be inconsistent. This is exactly where a least squares formulation of the problem will be useful, and the resulting rankings will impose an "approximately transitive" relationship.

Let's consider our previous example involving Teams A, B, and C, and include the new result that Team A defeats Team C by 3 points. Then, in order to impose a

ranking with a transitive relationship, we would impose the system of equations

$$\begin{aligned} m_A - m_B &= 10 \\ m_B - m_C &= 5 \\ m_A - m_C &= 3 \end{aligned}$$

on the ranking vector m, where m_A, m_B, and m_C all represent the ranking of the respective teams A, B, and C. In matrix form, this is just

$$\underbrace{\begin{bmatrix} 1 & -1 & 0 \\ 0 & 1 & -1 \\ 1 & 0 & -1 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} m_A \\ m_B \\ m_C \end{bmatrix}}_{m} = \underbrace{\begin{bmatrix} 10 \\ 5 \\ 3 \end{bmatrix}}_{b}$$

with $A \in \mathbb{R}^{g \times N}$, $m \in \mathbb{R}^N$, and $b \in \mathbb{R}^g$, where g and N represent the number of games played and teams, respectively. In our example, $N = g = 3$ so that A is square, but this is not generally the case.

Notice that $\text{rank}(A) = 2$ since the sum of the three columns produces the zero vector. Due to this, Theorem 5.14 implies that there are infinitely many least squares solutions, and this presents a bit of a problem since we would like to select just one to use for the ranking. More specifically, the normal equations are, of course, $A^T Am = A^T b$ where

$$A^T A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ -1 & -1 & 2 \end{bmatrix} \qquad \text{and} \qquad A^T b = \begin{bmatrix} 13 \\ -5 \\ -8 \end{bmatrix}.$$

In order to select a specific least squares solution, we will alter the last equation in the normal equations so that the resulting linear system is not singular, and thus possesses a unique solution that will define the respective ranking of each team. Note that we will alter the matrix $A^T A$ and the vector $A^T b$ and NOT the matrix A or vector b because m will be determined as a solution of $A^T Am = A^T b$ and will NOT satisfy the original linear system $Am = b$. The constraint that we choose for this alteration is a normalization of the rankings which enforces that they all must sum to zero, namely

$$m_A + m_B + m_C = 0.$$

Hence, the altered least squares matrix and vector become

$$A^T A = \begin{bmatrix} 2 & -1 & -1 \\ -1 & 2 & -1 \\ 1 & 1 & 1 \end{bmatrix} \qquad \text{and} \qquad b = \begin{bmatrix} 13 \\ -5 \\ 0 \end{bmatrix}.$$

Due to the structure of the matrix A - namely that each row contains exactly a single 1 entry, a single -1 entry and only zero entries elsewhere - imposing this constraint on $A^T A$ is guaranteed to increase the rank of $A^T A$. Additionally, removing the last row only removes redundant information about the system. So, we haven't lost any information in doing this, and we now have a normalized (in the sense that the sum of all rankings is zero), unique solution to the approximate

RMAC Team	Abbreviation	Record	Win %	Massey
CSU Pueblo	P	5 - 0	1.000	?
Colorado Mesa	CMU	5 - 1	0.833	?
Chadron State	C	3 - 2	0.600	?
Mines	CSM	3 - 2	0.600	?
Dixie State	D	3 - 2	0.600	?
Adams State	A	3 - 3	0.500	?
Black Hills	B	3 - 3	0.500	?
SD Mines	S	2 - 3	0.400	?
Fort Lewis	F	2 - 4	0.333	?
New Mexico Highlands	NM	1 - 4	0.200	?
Western State	W	0 - 6	0.000	?

Table 6.1: The eleven RMAC teams and their relative standings based on W-L record, winning percentage, and Massey rating from in-conference games.

ranking problem. In this case, we find the solution vector

$$m = \begin{bmatrix} 13/3 \\ -5/3 \\ -8/3 \end{bmatrix}.$$

These results imply that if Teams A and B played, we would expect Team A to win by 6 points since

$$m_A - m_B = \frac{13}{3} - \frac{-5}{3} = 6.$$

Similarly, we would expect Team A to defeat Team C by $m_A - m_C = 7$ points, and Team B to defeat Team C by $m_B - m_C = 1$ point. Finally, we note that QR factorization cannot be used to solve the system because we have altered A^TA directly and not via the matrix A.

Of course, this ranking method can be generalized to many more than three teams and can allow for teams to play multiple times (in which case, the total point differential in all games played by these teams is used) or for some teams to not play others at all. In general, we need not have the same number of games as teams; in fact, given N teams, a total of $g = \binom{N}{2} = \frac{N(N-1)}{2}$ games are possible should each team play every other team exactly once, and as previously mentioned, the $A \in \mathbb{R}^{g \times N}$ matrix is usually not square.

Example 46. As a final example, we might consider all eleven teams in the Rocky Mountain Athletic Conference (RMAC) and the outcomes of the games these teams have played within the conference (i.e., only against one another). For the sake of convenience, let's consider the $2017 - 2018$ season and the following table of the first 30 game results of that season:

Game	Winner	Loser	Point Diff.	Game	Winner	Loser	Point Diff.
1	P	CSM	31	16	B	S	1
2	CMU	W	26	17	CSM	W	38
3	NM	D	1	18	P	NM	43
4	C	F	6	19	C	A	47
5	B	A	6	20	CMU	F	26
6	CMU	C	14	21	B	NM	10
7	**CSM**	**NM**	**70**[1]	22	CSM	F	34
8	F	B	4	23	P	CMU	6
9	A	W	37	24	D	W	3
10	S	D	21	25	A	S	8
11	A	F	7	26	S	NM	38
12	C	W	42	27	F	W	1
13	CMU	CSM	2	28	CMU	B	12
14	D	B	22	29	D	C	14
15	P	S	20	30	P	A	46

Table 6.2: Results from the first 30 conference games among the 11 RMAC teams.

Inputting these results into Matlab, we can create a simple program to compute the Massey rating of each team, as outlined above:

[1]Mines beat New Mexico Highlands by a score of $70 - 0$ on September 9, 2017!

```
% Massey Example:
% Team abbreviation, alphabetically ordered
% A, B, C, CMU, CSM, D, F, NM, P, S,  W
% 1, 2, 3,  4,   5,  6, 7,  8, 9, 10, 11
clear;clc;

numg=30;
numt=11;
gamenum = 1:30;

win = [9, 4, 8, 3, 2, 4, 5, 7, 1, 10, 1, 3, 4, 6, 9, 2, 5, 9, 3,...
    4, 2, 5, 9, 6, 1, 10, 7, 4, 6, 9];

loss =[5, 11, 6, 7, 1, 3, 8, 2, 11, 6, 7, 11, 5, 2, 10, 10, 11, ...
    8, 1,...
    7, 8, 7, 4, 11, 10, 8, 11, 2, 3, 1];

diff = [31, 26, 1, 6, 6, 14, 70, 4, 37, 21, 7, 42, 2, 22, 20, 1, ...
    38,...
    43, 47, 26, 10, 34, 6, 3, 8, 38, 1, 12, 14, 46]';


Aw = sparse(gamenum, win, 1,numg,numt);
nnz(Aw);
full(Aw);

Al = sparse(gamenum, loss,-1,numg,numt);
nnz(Al);
full(Al);

sum(full(Aw))
sum(full(Al))


A = full(Al) + full(Aw)
sum(A')
%%

% Compute projection onto Col(A)
M = A'*A;
rank(M)
size(M)
c= A'*diff;

if rank(M) < size(M, 1)
    %Augment M to build Massey matrix
    M(end, :) = ones(1, numt);
    c(end) = 0;
end

r = M\c;

[finalrank, index] = sort(r, 'descend')
```

Rank	Team	W-L%
1	CSU Pueblo (P)	1.000
2	CO Mesa (CMU)	0.833
3	Chadron (C)	0.600
4	Mines (CSM)	0.600
5	Dixie (D)	0.600
6	Adams (A)	0.500
7	Black Hills (B)	0.500
8	SD Mines (S)	0.400
9	Fort Lewis (F)	0.333
10	NM Highlands (NM)	0.200
11	Western (W)	0.000

Table 6.3: RMAC teams ranked by winning percentage.

Rank	Team	Massey
1	CSU Pueblo (P)	30.56
2	Mines (CSM)	19.88
3	CO Mesa (CMU)	15.37
4	Chadron (C)	7.61
5	SD Mines (S)	3.12
6	Dixie (D)	-5.14
7	Adams (A)	-7.59
8	Black Hills (B)	-8.26
9	Fort Lewis (F)	-10.37
10	Western (W)	-21.21
11	NM Highlands (NM)	-23.97

Table 6.4: RMAC teams ranked by Massey index.

Using this code, we run the approximate least squares algorithm to include the augmentation of the A^TA matrix and implement Massey's idea. The results provide a reordering of the previous standings, which were based solely on win percentage. Instead, the Massey index provides an expected margin of victory for each team under the assumption that they play an average team, i.e. a team with a Massey rating of zero.

We can use this information to actually predict scores of future games by merely subtracting the respective Massey indices of the teams involved in an upcoming game. For instance, the previous game results and the Massey ratings tell us that should the Colorado School of Mines (CSM) and South Dakota School of Mines (S) play, we would expect the CSM team to win (as they have a greater Massey rating) and their expected margin of victory would be $19.88 - 3.12 = 16.76$ points. Of course, some teams possess negative Massey ratings - they must since we've imposed that the sum of all the ratings is zero - and thus when a top team plays a lower-ranked team, the expected margin of victory can be increasingly wide. For instance, if CSU Pueblo played NM Highlands, we would expect the former team to win by $30.56 - (-23.97) = 54.53$ points.

Finally, we note that the Massey rating (Table 6.4) changes a large amount of the traditional ranking structure in Table 6.3. As an example, notice that Western State is winless, but has a greater Massey index than a one-win NM Highlands team. This is due to the emphasis the Massey algorithm places on point differential and some lopsided defeats that Highlands suffered at the beginning of the season, including a $70 - 0$ loss to CSM.

For more information on (deterministic) ranking algorithms, see [8, 13].

Exercises - Linear Regression & Ranking

For problems which require computational simulation, please print and submit both your code and results (e.g., pictures).

Problem 6.1. Consider the following Olympic Gold Medal Winning times (in seconds) for the Men's 100 Meter Dash:

Year	Time	Year	Time
1960	10.32	1988	9.92
1964	10.06	1992	9.96
1968	9.95	1996	9.84
1972	10.14	2000	9.87
1976	10.06	2004	9.85
1980	10.25	2008	9.69
1984	9.99	2012	9.63

(a) Use Least Squares fitting to compute the line $T = \alpha_0 + \alpha_1 x$ and the parabola $T = \beta_0 + \beta_1 x + \beta_2 x^2$ which best fit the data, where T represents the 100 Meter time, y is the true Olympic year, and x represents the scaled Olympic year given by

$$x = \frac{y - 1960}{52}.$$

Include your Matlab code and output for the models.

(b) Graph the data and your first Least Squares curve from part (a) on the same axes, and then do this again with your second Least Squares curve from part (a)

(c) Using each of your resulting models, predict the gold medal time for the 2020 Olympics.

Problem 6.2. Consider the following match results for five different football teams, denoted by Teams 1 through 5:

Winner	Loser	Difference	Winner	Loser	Difference
1	2	7	3	4	4
3	5	1	2	4	10
1	5	20	3	2	12
2	5	7	3	1	3
4	1	24	5	4	1

(a) Determine the win-loss records of each team and their corresponding winning percentage, and then rank them by winning percentage.

(b) Compute the Massey indices of these teams and their associated rankings.

(c) Change the result in the second portion of the last row from $5 - 4 - 1$ to $5 - 4 - 20$, and recompute the indices and rankings. Does this elevate Team 5 in the standings? How large must the strength of victory (a positive integer) be for Team 5 in this last game in order to boost them to the top of the rankings?

Chapter 7

Operator Decompositions and Factorizations

We will discuss a number of different categorizations of linear operators related to their eigenvalue/eigenvector properties, and focus on their implications to analogous matrices with these same properties. In a first course in Linear Algebra, you may be introduced to the notion of an eigenvalue using the determinant of a given matrix. Here, we would like to remove the emphasis on the determinant since this operation is not as useful for general linear operators on (infinite-dimensional) vector spaces. Additionally, for computationally-oriented problems it is very rare to compute the determinant of a matrix (much like it is rare in practice to actually invert a matrix), and the algorithm we typically employ by hand to find a determinant is extraordinarily inefficient for a computer. For all of these reasons, we will not utilize determinants going forward. Finally, we will generally assume throughout that the vector space $\mathcal{V}$ is a Hilbert space with inner product $\langle \cdot, \cdot \rangle$, though many of the ideas of Spectral Theory (concerning eigenvalues and eigenvectors) can be formulated without an inner product, and we will stipulate weaker assumptions on $\mathcal{V}$ when possible.

7.1 Introduction

We begin by defining some familiar concepts in the generalized framework of a vector space, including eigenvalues and eigenvectors of linear operators.

Definition 7.1. For any vector space $\mathcal{V}$, define the **identity** operator $\mathbb{I} : \mathcal{V} \to \mathcal{V}$ by $I(v) = v$ for all $v \in \mathcal{V}$.

Definition 7.2. Let $T : \mathcal{V} \to \mathcal{V}$ be bounded and linear. Then,

1. A scalar $\lambda \in \mathbb{C}$ is an **eigenvalue** of T if there is $v \in \mathcal{V}$ with $v \neq 0$ such that

$$T(v) = \lambda v. \tag{7.1}$$

   Associated to this eigenvalue λ, we call any $v \in \mathcal{V} \setminus \{0\}$ satisfying (7.1) an **eigenvector**. Note that $\lambda \in \mathbb{C}$, not necessarily $\lambda \in \mathbb{K}$.

2. The (point) **spectrum** of T is the set of all eigenvalues, namely

$$\sigma(T) = \Big\{ \lambda \in \mathbb{C} : T(v) = \lambda v \text{ for some } v \in \mathcal{V} \setminus \{0\} \Big\}.$$

3. For every $\lambda \in \sigma(T)$, we define the linear operator $R_\lambda : \mathcal{V} \to \mathcal{V}$ by

$$R_\lambda(v) = (T - \lambda I)v,$$

referred to as the (inverse) **resolvent operator**, and the subspace

$$\mathcal{E}_\lambda = \mathrm{Ker}(R_\lambda) \subseteq \mathcal{V}$$

called the **eigenspace** (or collection of all eigenvectors) of T associated to λ.

4. The **spectral radius** of T is defined by

$$\rho(T) = \sup\{|\lambda| : \lambda \in \sigma(T)\}.$$

Comment. For $\dim(\mathcal{V}) < \infty$, it is equivalent to state that $\lambda \in \mathbb{C}$ is an eigenvalue of T if and only if R_λ is not one-to-one.

Example 47. We have to be a bit careful with eigenvalues. For instance, eigenvalues do not always exist in $\mathbb{K}$ for a given T. Consider $\mathcal{V} = \mathbb{R}^2$ and

$$T(v) = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} v.$$

Then, (7.1) implies that any eigenvalue λ satisfies

$$\lambda^2 + 1 = 0.$$

Thus, $\lambda = \pm i$ but these values are not in $\mathbb{K} = \mathbb{R}$. Of course, this can be remedied by choosing $\mathcal{V} = \mathbb{C}^2$ and considering the same mapping T with domain $\mathbb{C}^2$. In fact, this is implied by the next theorem.

Theorem 7.1. Let $\mathcal{V}$ be a Hilbert space and $T : \mathcal{V} \to \mathcal{V}$ be bounded and linear. Then, there exists $\lambda \in \sigma(T)$ such that $|\lambda| = \rho(T)$.

The proof of this result uses tools of Complex Analysis beyond the scope of this class, but see [9] for more details.

Before discussing special classes of linear operators, we will need one more general result concerning eigenvalues and eigenvectors, which you may have seen in a first course in Linear Algebra.

Lemma 7.2. Let $T : \mathcal{V} \to \mathcal{V}$ be linear and $k \in \mathbb{N}$. If $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of T and $v_1, \ldots, v_k$ are any associated eigenvectors, then the set $S = \{v_1, \ldots, v_k\}$ is linearly independent.

Proof. The proof is assigned as a homework problem (cf. Problem 7.3). □

7.2 Diagonalizable Operators and Similar Matrices

Definition 7.3. Let $T : \mathcal{V} \to \mathcal{V}$ be linear. We say T is **diagonalizable** if there exists a basis for $\mathcal{V}$, say $\{v_i\}_{i\in I}$, such that v_i is an eigenvector of T for every $i \in I$.

Diagonalizable operators are interesting and important because their operation inherently captures the entire vector space on which they're defined, namely through their associated eigenvectors. As a special case $\mathcal{V} = \mathbb{C}^p$ and $T(v) = Av$ for some $A \in \mathbb{C}^{p\times p}$ has the same definition as above, but these linear operators can be characterized by a specific matrix factorization.

Theorem 7.3. Let $p \in \mathbb{N}$ and $A \in \mathbb{C}^{p\times p}$ be given and define $T : \mathbb{C}^p \to \mathbb{C}^p$ by $T(v) = Av$ for every $v \in \mathbb{C}^p$. Then, T (or A) is diagonalizable if and only if there is a nonsingular matrix $P \in \mathbb{C}^{p\times p}$ and a diagonal matrix $\Lambda \in \mathbb{C}^{p\times p}$ such that

$$A = P\Lambda P^{-1}.$$

Proof. The proof is left as a homework exercise (cf. Problem 7.4), but the idea of the proof is as follows. Because P is nonsingular, we can multiply by P on the right of the factorization $A = P\Lambda P^{-1}$ so that it is equivalent to writing $AP = P\Lambda$. From this representation, we could write each column of AP as Av_j where each v_j is the jth column of P. Since Λ is diagonal, each column of $P\Lambda$ can then be written as $\lambda_j v_j$ where λ_j is the jth diagonal entry of Λ. With the similarity of the resulting equality to (7.1), we would expect eigenvalues and eigenvectors to be involved in constructing the matrices P and Λ. □

Comment. The most crucial idea taken from this theorem is that **within any matrix diagonalization (i.e. $A = P\Lambda P^{-1}$), the diagonal matrix Λ contains the eigenvalues of A and the matrix P contains the eigenvectors of A.** We will study many such diagonalizations and all involve special properties of eigenvalues and eigenvectors of A.

Corollary 7.1. If $A \in \mathbb{C}^{p\times p}$ possesses p distinct eigenvalues, then A is diagonalizable (or equivalently, $T(v) = Av$ is diagonalizable).

Proof. Because A has p distinct eigenvalues $\lambda_1, \ldots, \lambda_p$, the set of corresponding eigenvectors $S = \{v_1, \ldots, v_p\}$ is linearly independent by Lemma 7.2, where

$$T(v_k) = \lambda_k v_k, \qquad \text{i.e.} \quad Av_k = \lambda_k v_k$$

for $k = 1, ..., p$. Since S contains p linearly independent vectors in $\mathbb{C}^p$, they must also form a basis for $\mathbb{C}^p$. As S consists only of eigenvectors of A, the matrix A is diagonalizable by definition. □

Example 48. We provide an example of both a diagonalizable and non-diagonalizable matrix, which are nearly identical save for one entry. Define the matrices

$$A_1 = \begin{bmatrix} 2 & -1 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix}, \qquad A_2 = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 2 & 1 \\ 0 & 0 & 3 \end{bmatrix}.$$

Because A_1 and A_2 are **upper triangular** matrices, their eigenvalues are exactly their diagonal entries, namely $\lambda_1 = 2$ (with algebraic multiplicity 2) and $\lambda_2 = 3$. Computing the eigenvectors for A_1, we find

$$\lambda_1 = 2 \to \text{Nul}(A_1 - 2I) = \text{Nul}\left(\begin{bmatrix} 0 & -1 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}\right) = \left\{ \alpha \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} : \alpha \in \mathbb{R} \right\}$$

$$\lambda_2 = 3 \to \text{Nul}(A_1 - 3I) = \text{Nul}\left(\begin{bmatrix} -1 & -1 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}\right) = \left\{ \beta \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix} : \beta \in \mathbb{R} \right\}$$

Thus, no set of eigenvectors of A_1 form a basis for $\mathbb{R}^3$, and A_1 is not diagonalizable!

Similarly, computing the eigenvectors for A_2, we find

$$\lambda_1 = 2 \to \text{Nul}(A_2 - 2I) = \text{Nul}\left(\begin{bmatrix} 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{bmatrix}\right) = \left\{ \alpha \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} + \beta \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} : \alpha, \beta \in \mathbb{R} \right\}$$

$$\lambda_1 = 3 \to \text{Nul}(A_2 - 3I) = \text{Nul}\left(\begin{bmatrix} -1 & 0 & 1 \\ 0 & -1 & 1 \\ 0 & 0 & 0 \end{bmatrix}\right) = \left\{ \gamma \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} : \gamma \in \mathbb{R} \right\}$$

Because the set of vectors

$$B = \left\{ \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

forms a basis for $\mathbb{R}^3$, it follows that A_2 is diagonalizable. In particular, letting

$$P = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \qquad \text{and} \qquad \Lambda = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 3 \end{bmatrix}$$

we find

$$A_2 = P\Lambda P^{-1}.$$

Definition 7.4. Given $A, B \in \mathbb{C}^{p\times p}$, we say A and B are **similar** (often denoted by $A \sim B$) if there is a nonsingular $P \in \mathbb{C}^{p\times p}$ such that

$$A = PBP^{-1}$$

or equivalently, $B = P^{-1}AP$.

With this definition, a result immediately arises by merging the notion of matrix similarity with the equivalence theorem regarding diagonalizability.

Corollary 7.2. A matrix $A \in \mathbb{C}^{p\times p}$ is diagonalizable if and only if A is similar to a diagonal matrix.

7.3 Jordan Form

With these results in place, we see that not all square matrices are similar to a diagonal matrix, which is unfortunate because given an arbitrary square matrix A, it would often be much easier to work with an associated diagonal matrix Λ to perform some computations, and then merely transform, via the nonsingular matrix P, back to A. Since some matrices are not similar to a diagonal matrix, it's useful to know for matrices without this property, what is the simplest matrix to which they are similar? The short answer is that there is a simpler form for every matrix, called the **Jordan form** of A, but as we will see, it is not quite as friendly as a diagonal matrix. Before we can define the Jordan form of a matrix, however, we will first need some other basic terms.

Definition 7.5. A matrix $A \in \mathbb{R}^{p\times p}$ is called **block diagonal** if it is of the form

$$A = \begin{bmatrix} A_1 & 0 & \cdots & 0 \\ 0 & A_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & A_k \end{bmatrix}$$

where A_j is a square matrix for every $j = 1, \ldots, k$ with the diagonal of A_j lying on the diagonal of A, and 0 represents an appropriately sized matrix of zeros. It is important to note that no A_j need be a diagonal matrix in this representation.

Definition 7.6. A **Jordan block** with value λ is a square, upper-triangular matrix whose diagonal entries are λ, whose super diagonal entries (i.e. the entries immediately above the diagonal) are all 1, and whose remaining entries are 0.

Example 49. The following are general forms (for $\lambda \in \mathbb{R}$) of Jordan blocks of size 1×1, 2×2, and 3×3, respectively:

$$J_1(\lambda) = \begin{bmatrix} \lambda \end{bmatrix}, \qquad J_2(\lambda) = \begin{bmatrix} \lambda & 1 \\ 0 & \lambda \end{bmatrix}, \qquad J_3(\lambda) = \begin{bmatrix} \lambda & 1 & 0 \\ 0 & \lambda & 1 \\ 0 & 0 & \lambda \end{bmatrix}.$$

Definition 7.7. A **Jordan form matrix** is a block diagonal matrix in which all of the (non-zero) blocks are Jordan blocks.

Example 50. Consider the 6×6 matrix A defined by

$$A = \begin{bmatrix} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ 0 & 0 & -3 & 1 & 0 & 0 \\ 0 & 0 & 0 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 & -3 & 0 \\ 0 & 0 & 0 & 0 & 0 & -1 \end{bmatrix}.$$

Then, partitioning submatrices accordingly, we can express this as

$$A = \left[\begin{array}{cc|cccc} 2 & 1 & 0 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & -3 & 1 & 0 & 0 \\ 0 & 0 & 0 & -3 & 1 & 0 \\ 0 & 0 & 0 & 0 & -3 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & -1 \end{array}\right]$$

which is of the form

$$A = \left[\begin{array}{cc|ccc|c} \lambda_1 & 1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_1 & 0 & 0 & 0 & 0 \\ \hline 0 & 0 & \lambda_2 & 1 & 0 & 0 \\ 0 & 0 & 0 & \lambda_2 & 1 & 0 \\ 0 & 0 & 0 & 0 & \lambda_2 & 0 \\ \hline 0 & 0 & 0 & 0 & 0 & \lambda_3 \end{array}\right] = \left[\begin{array}{c|c|c} J_2(\lambda_1) & 0 & 0 \\ \hline 0 & J_3(\lambda_2) & 0 \\ \hline 0 & 0 & J_1(\lambda_3) \end{array}\right]$$

where $\lambda_1 = 2$, $\lambda_2 = 3$, $\lambda_3 = -1$, and $J_k(\lambda)$ is the Jordan block of size $k \times k$, as in the previous example. Therefore, with this representation in terms of the Jordan blocks of A, we see that it is a Jordan form matrix.

Now that we know what a Jordan form matrix looks like, we can state the main result, namely that every square matrix can be put into this special form via a similarity transformation.

Theorem 7.4. Let $A \in \mathbb{C}^{p\times p}$ be given. Then, there is a Jordan form matrix $J \in \mathbb{C}^{p\times p}$ such that A and J are similar.

Though we will not prove this result, it's worthwhile to at least discuss the main idea. Recall that diagonalizable matrices are decomposed into their respective eigenvalues and eigenvectors, where the eigenvectors form a linearly independent set. However, what happens if the eigenvectors (corresponding to the same eigenvalue) do not form a linearly independent set? In other words, what if the geometric multiplicity is strictly less than the algebraic multiplicity of a particular eigenvalue? Clearly the structure of the Λ matrix in an attempted diagonalization would be altered, and this is exactly what leads to the Jordan form, rather than the diagonal form. This is perhaps best illustrated by the following example.

Example 51. Consider the following matrices:

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \qquad \text{and} \qquad B = \begin{bmatrix} 3 & 1 \\ -1 & 5 \end{bmatrix}.$$

Clearly A is already diagonal, and thus its diagonalization consists of $\Lambda = A$ and any 2×2 matrix P with linearly independent columns since

$$P\Lambda P^{-1} = P4\mathbb{I}P^{-1} = 4PP^{-1} = 4\mathbb{I} = A.$$

This occurs regardless of the invertible matrix P chosen to represent the eigenvectors. Additionally, notice that A has only one eigenvalue $\lambda = 4$, which has algebraic multiplicity two and geometric multiplicity two.

Unfortunately, B appears to be quite different. However, computing its eigenvalues, we find $\lambda = 4$ is also the only eigenvalue of B, as was the case for A. Thus, if we attempt to diagonalize B, we find $\Lambda = 4\mathbb{I}$ is the appropriate matrix for the eigenvalues. This attempt falls apart in computing the eigenvectors of B, as we find

$$\mathcal{E}_4 = \left\{ \alpha \begin{bmatrix} 1 \\ 1 \end{bmatrix} : \alpha \in \mathbb{R} \right\}$$

is the eigenspace corresponding to $\lambda = 4$. In particular, $\dim(\mathcal{E}_4) = 1$ and therefore we cannot create a basis for $\mathbb{R}^2$ merely from eigenvectors of B. However, if we alter $\Lambda = 4\mathbb{I}$ slightly by using a Jordan block, namely

$$\tilde{\Lambda} = J_2(4) = \begin{bmatrix} 4 & 1 \\ 0 & 4 \end{bmatrix},$$

then solving the matrix equation $P\tilde{\Lambda} = BP$ or

$$\begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \begin{bmatrix} 4 & 1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ -1 & 5 \end{bmatrix} \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix}$$

provides the solution

$$p_{11} = p_{21}, \qquad p_{11} + p_{12} = p_{22}.$$

Hence, any matrix of the form

$$P = s \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} + t \begin{bmatrix} 0 & 1 \\ 0 & 1 \end{bmatrix}$$

for $s, t \in \mathbb{R}$ will satisfy the diagonalization relationship. Finally, we can simplify this by choosing $s = 1$ and $t = 0$ so that

$$P = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \qquad \text{and} \qquad P^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & 1 \end{bmatrix}$$

which yields $P\tilde{\Lambda}P^{-1} = B$.

Comment. Though the Jordan form of a square matrix is not as common or easy to use as a diagonal matrix, they do arise within a variety of mathematical fields, including

1. Ordinary Differential Equations - in computing the general solution of linear systems
2. Complex Analysis - in proving invariant subspace decompositions of $\mathbb{C}^n$
3. Algebra - in proving the Cayley-Hamilton Theorem (i.e., $p(A) = 0$)

However, it should be noted that the Jordan form is not used computationally due to numerical instabilities. Instead, a slightly different decomposition called the **Schur Form** is used, and we will discuss this in the next section.

7.4 Unitary operators and the Schur Form

Definition 7.8. Let $\mathcal{V}$ and $\mathcal{W}$ be Hilbert spaces with $T : \mathcal{V} \to \mathcal{W}$ bounded and linear. Then, we say T is **unitary** if

$$T^*T = \mathbb{I}_{\mathcal{V}} \qquad \text{and} \qquad TT^* = \mathbb{I}_{\mathcal{W}},$$

i.e. $T^*(T(v)) = v$ for every $v \in \mathcal{V}$ and $T(T^*(w)) = w$ for every $w \in \mathcal{W}$.

Comment. For $U \in \mathbb{C}^{p\times p}$ this definition is equivalent to

$$U^H U = UU^H = \mathbb{I}_p$$

and these are called **unitary matrices**. Equivalently, $U \in \mathbb{C}^{p\times p}$ is unitary if and only if U is nonsingular and $U^{-1} = U^H$.

For $Q \in \mathbb{R}^{p\times p}$ this definition is equivalent to

$$Q^T Q = QQ^T = \mathbb{I}_p$$

and these are called **orthogonal matrices**. Equivalently, $Q \in \mathbb{R}^{p\times p}$ is orthogonal if and only if Q is nonsingular and $Q^{-1} = Q^T$.

Example 52. Here are a few examples of familiar unitary operators:

1. The matrix $A = \begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} \\ -1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$ is an orthogonal matrix.

2. The matrix $A = \frac{1}{2}\begin{bmatrix} 1+i & 1-i \\ 1-i & 1+i \end{bmatrix}$ is a unitary matrix.

3. Let $\mathscr{F} : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ be the Fourier transform, defined by

$$\mathscr{F}[u](\xi) = \frac{1}{\sqrt{2\pi}} \int e^{-ix\xi} u(x) dx.$$

   Then, $\mathscr{F}$ is a unitary operator, i.e. $\mathscr{F}^*[\mathscr{F}[u]](x) = u(x)$ for every $u \in L^2(\mathbb{R})$. In Harmonic Analysis this implies the **Plancherel Theorem**, which states

$$\|u\|_{L^2(\mathbb{R})} = \|\mathscr{F}[u]\|_{L^2(\mathbb{R})}.$$

   As you may have seen previously, the Fourier Transform is also a powerful tool used to solve a variety of linear PDEs.

Next, we investigate some properties of unitary operators. In particular, they preserve angles and norms of their arguments.

Theorem 7.5. Let $T : \mathcal{V} \to \mathcal{W}$ be unitary. Then,

1. For every $u, v \in \mathcal{V}$, we have

$$\langle T(u), T(v)\rangle_{\mathcal{W}} = \langle u, v\rangle_{\mathcal{V}}.$$

2. For every $v \in \mathcal{V}$, we have

$$\|T(v)\|_{\mathcal{W}} = \|v\|_{\mathcal{V}}.$$

3. For every $u, v \in \mathcal{V}$, we have

$$\|T(u) - T(v)\|_{\mathcal{W}} = \|u - v\|_{\mathcal{V}}.$$

Proof. We prove the results sequentially and use each previous conclusion to establish the next one.

To prove the first result, we merely use the definition of the adjoint and the unitary property of T to find

$$\langle T(u), T(v)\rangle_{\mathcal{W}} = \langle u, T^*(T(v))\rangle_{\mathcal{V}} = \langle u, v\rangle_{\mathcal{V}}$$

for any $u, v \in \mathcal{V}$.

With this, the second property follows by using the first result so that

$$\|T(v)\|_{\mathcal{W}}^2 = \langle T(v), T(v)\rangle_{\mathcal{W}} = \langle v, v\rangle_{\mathcal{V}} = \|v\|_{\mathcal{V}}^2$$

for every $v \in \mathcal{V}$. Taking the square root of both sides then yields the conclusion. Finally, to prove the last conclusion we use the linearity of T and the second result, to find

$$\|T(u) - T(v)\|_{\mathcal{W}} = \|T(u - v)\|_{\mathcal{W}} = \|u - v\|_{\mathcal{V}}$$

for every $u, v \in \mathcal{V}$, and the proof is complete. □

Comment. We note a few implications of this theorem for matrices.

1. The first property in this theorem states that T is a **conformal** map, i.e. an angle-preserving transformation, while second property states that T is an **isometry**, i.e. a distance-preserving transformation.

2. Note that each of the conclusions of Theorem 7.5 holds for linear operators $T(x) = Ux$ where $U \in \mathbb{C}^{p\times p}$ or $T(x) = Qx$ where $Q \in \mathbb{R}^{p\times p}$. More specifically, the norm within these conclusions is the two-norm induced by the inner product; for instance, the second conclusion becomes

   $$\|Ux\|_2 = \|x\|_2$$

   for every $x \in \mathbb{C}^p$.

3. For $U \in \mathbb{C}^{p\times p}$ or $Q \in \mathbb{R}^{p\times p}$, the second result of the theorem further implies

   $$\|U\|_2 = \|Q\|_2 = 1$$

   because, for instance

   $$\|U\|_2 = \max_{x\in\mathbb{C}^p\setminus\{0\}} \frac{\|Ux\|_2}{\|x\|_2} = \max_{x\in\mathbb{C}^p\setminus\{0\}} 1 = 1.$$

Theorem 7.6. Let $U \in \mathbb{C}^{p\times p}$ and $Q \in \mathbb{R}^{p\times p}$ be given. Then,

1. U is unitary if and only if the columns of U are orthonormal with respect to the standard inner product on $\mathbb{C}^p$, namely $\langle x, y\rangle_{\mathbb{C}^p} = x^H y$.

2. Q is orthogonal if and only if the columns of Q are orthonormal with respect to the standard inner product on $\mathbb{R}^p$, namely $\langle x, y\rangle_{\mathbb{R}^p} = x^T y$.

Proof. These results are very similar, and the proof of the second is included as a homework exercise (cf. Problem 7.5). The main idea is to use the fact that the entries of Q^TQ are merely inner products of columns of Q. Hence, the orthonormality of these columns is equivalent to these entries being 1 along the diagonal and 0 elsewhere. □

Theorem 7.7. Let $P, R \in \mathbb{C}^{p\times p}$ be given. Then,

1. The product PR is unitary if P and R are unitary.
2. P is unitary if and only if P^H is unitary.
3. If P and R are real-valued, then the product PR is orthogonal if P and R are orthogonal.
4. If P is real-valued, then P is orthogonal if and only if P^T is orthogonal.

Proof. The proof of the first and third results are similar and the former is a homework exercise (cf. Problem 7.7). The second and fourth results are also very similar, so we merely prove the former.
Notice that P unitary is equivalent to $P^HP = PP^H = \mathbb{I}$. Additionally, P^H unitary means exactly that

$$(P^H)^HP^H = \mathbb{I} \qquad \text{and} \qquad P^H(P^H)^H = \mathbb{I},$$

or simplifying these expressions

$$PP^H = \mathbb{I} \qquad \text{and} \qquad P^HP = \mathbb{I},$$

which are merely the above equivalence properties for P unitary. □

Definition 7.9. Let $A, B \in \mathbb{C}^{p\times p}$ be given. We say A and B are **unitarily similar** if there exists a unitary $U \in \mathbb{C}^{p\times p}$ such that

$$A = UBU^H.$$

Similarly, we say $A, B \in \mathbb{R}^{p\times p}$ are **orthogonally similar** if there is an orthogonal $Q \in \mathbb{R}^{p\times p}$ such that

$$A = QBQ^T.$$

Comment. Because unitary and orthogonal matrices are necessarily invertible, this is just a special case of similar matrices $A \sim B$. Hence, all properties of similar matrices hold for these as well.

Furthermore, notice that if $A = UBU^H$ where $U^{-1} = U^H$, then multiplying on the left and right side of the similarity equation yields $B = U^{-1}A(U^H)^{-1} = U^HAU$. This is just the statement that B is unitarily similar to A with similarity matrix U^H, rather than A is unitarily similar to B with similarity matrix U, which we can now see, are equivalent statements.

Defining these special notions of similarity allows us to introduce the Schur form of a matrix, which we briefly hinted at earlier in our discussion of the Jordan form.

Theorem 7.8. Let $A \in \mathbb{C}^{p\times p}$ be given. Then, the following statements hold.

1. There is a unitary matrix $U \in \mathbb{C}^{p\times p}$ and an upper triangular $T \in \mathbb{C}^{p\times p}$ such that
$$A = UTU^H.$$

2. If $A \in \mathbb{R}^{p\times p}$ has real eigenvalues, then there is an orthogonal $Q \in \mathbb{R}^{p\times p}$ that provides the analogous decomposition
$$A = QTQ^T.$$
Thus, every square matrix A is unitarily similar (or orthogonally similar, if A is real-valued with real eigenvalues) to an upper triangular matrix. The matrix T is referred to as the **Schur decomposition** or **Schur form** of A.

Comment. We remark that the diagonal entries of T will merely be the eigenvalues of A repeated according to their algebraic multiplicity. Additionally, U can be chosen so that these eigenvalues appear in any order desired. Hence, the Schur form is not explicitly unique.

Instead of presenting a more abstract proof of this theorem, we will introduce an iterative algorithm that will construct the Schur form of a given matrix. Though this is done by way of a specific example, the algorithm will generalize for use with any square matrix.

Example 53. Define $A \in \mathbb{R}^{p\times p}$ by
$$A = \begin{bmatrix} 0.2 & 0.6 & 0 \\ 1.6 & -0.2 & 0 \\ -1.6 & 1.2 & 3 \end{bmatrix}.$$

We compute the Schur form using the following iterative algorithm.

1. Find an eigenvalue and corresponding unit eigenvector of A:

   For this particular matrix A, we find
$$\lambda_1 = 3 \qquad \text{and} \qquad v_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}.$$

2. Construct a unitary/orthogonal matrix from v_1:

   Since A is real-valued, we actually construct an orthogonal matrix, and as we have a choice, we choose a simple one. We let $q_1 = v_1$ and define the orthogonal matrix Q_1 by
$$Q_1 = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}.$$

3. Compute the current approximation to the unitary decomposition:

$$A' = Q_1^T A Q_1 = \begin{bmatrix} q_1^T \\ q_2^T \\ q_3^T \end{bmatrix} \left[\begin{array}{c|c|c} Aq_1 & Aq_2 & Aq_3 \end{array} \right].$$

Because $Aq_1 = \lambda_1 q_1$ in our example, we can simplify this representation to

$$\begin{aligned} A' &= \begin{bmatrix} q_1^T \\ q_2^T \\ q_3^T \end{bmatrix} \left[\begin{array}{c|c|c} 3q_1 & Aq_2 & Aq_3 \end{array} \right] \\ &= \begin{bmatrix} 3 & \alpha & \beta \\ 0 & & \\ 0 & & B \end{bmatrix} \end{aligned}$$

where

$$\begin{aligned} \alpha &= q_1^T A q_2 = 1.2 \\ \beta &= q_1^T A q_3 = -1.6 \end{aligned}$$

and

$$B = \begin{bmatrix} q_2^T A q_2 & q_2^T A q_3 \\ q_3^T A q_2 & q_3^T A q_3 \end{bmatrix} = \begin{bmatrix} -0.2 & 1.6 \\ 0.6 & 0.2 \end{bmatrix}.$$

4. Continue this iterative process on the remaining submatrix:

With one full iteration of the algorithm complete, we began with $A \in \mathbb{R}^{3\times 3}$, computed the first row and column in the associated Schur decomposition, and are now left with $B \in \mathbb{R}^{2\times 2}$. Hence, we repeat the procedure on the remaining submatrix B.

Computing an eigenvalue and unit eigenvector, we find

$$\lambda_2 = 1 \qquad \text{with} \qquad v_2 = \begin{bmatrix} 0.8 \\ 0.6 \end{bmatrix}.$$

Additionally, the remaining eigenvalue of B arises naturally, and it is $\lambda_3 = -1$.

Next, we construct an orthogonal matrix from v_2. In particular, we let $w_1 = v_2$ and define the orthogonal matrix W by

$$W = \begin{bmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{bmatrix}.$$

With this, we compute the unitary decomposition for B using the eigenpair relationship $Bv_2 = \lambda_2 v_2$ so that $Bw_1 = \lambda_2 w_1 = w_1$. Hence, we find

$$\begin{aligned} B' = W^T B W &= \begin{bmatrix} w_1^T \\ w_2^T \end{bmatrix} \left[\begin{array}{c|c} Bw_1 & Bw_2 \end{array} \right] \\ &= \begin{bmatrix} w_1^T \\ w_2^T \end{bmatrix} \left[\begin{array}{c|c} w_1 & Bw_2 \end{array} \right] \\ &= \begin{bmatrix} 1 & \gamma \\ 0 & \lambda_3 \end{bmatrix} \end{aligned}$$

where

$$\gamma = w_1^T B w_2 = 1 \qquad \text{and} \qquad \lambda_3 = -1.$$

Here, the condition $b'_{2,2} = \lambda_3 = -1$ occurs because the eigenvalues of B must lie on the diagonal of the constructed Schur matrix - notice, for instance, that $a'_{1,1} = \lambda_1$ and $b'_{1,1} = \lambda_2$.

5. Assemble all submatrix Schur decompositions:

So, we have an upper triangular matrix that is ALMOST the final product. We merely need to place the submatrix decomposition into a larger form, multiply the orthogonal decompositions, and use the fact that the product of orthogonal matrices remains orthogonal. In particular, let

$$Q_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & W \end{bmatrix}$$

and define

$$T = Q_2^T Q_1^T A Q_1 Q_2.$$

Since Q_1 and Q_2 are orthogonal, it follows that $Q = Q_1 Q_2$ must also be orthogonal by Theorem 7.7. Hence, T becomes

$$T = Q^T A Q$$

which can be inverted to arrive at

$$A = Q T Q^T.$$

Thus, A is unitarily similar to T, and we merely need to compute T. Doing this, we finally compute

$$\begin{aligned} T &= Q_2^T \underbrace{Q_1^T A Q_1}_{A'} Q_2 \\ &= Q_2^T A' Q_2 \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & W^T \end{bmatrix} \begin{bmatrix} \lambda_1 & \alpha & \beta \\ 0 & & \\ 0 & & B \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & W \end{bmatrix} \\ &= \begin{bmatrix} 1 & 0 & 0 \\ 0 & & \\ 0 & & W^T \end{bmatrix} \begin{bmatrix} \lambda_1 & [\alpha & \beta]W \\ 0 & & \\ 0 & & BW \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & [\alpha & \beta]W \\ 0 & & \\ 0 & & \underbrace{W^T B W}_{=B'} \end{bmatrix} \\ &= \begin{bmatrix} \lambda_1 & [\alpha & \beta]W \\ 0 & \lambda_2 & \gamma \\ 0 & 0 & \lambda_3 \end{bmatrix}. \end{aligned}$$

Since

$$\begin{bmatrix} \alpha & \beta \end{bmatrix} W = \begin{bmatrix} 1.2 & -1.6 \end{bmatrix} \begin{bmatrix} 0.8 & -0.6 \\ 0.6 & 0.8 \end{bmatrix} = \begin{bmatrix} 0 & -2 \end{bmatrix},$$

we find

$$T = \begin{bmatrix} 3 & 0 & -2 \\ 0 & 1 & 1 \\ 0 & 0 & -1 \end{bmatrix},$$

which is indeed upper triangular, and thus $A = QTQ^T$ where $Q = Q_1Q_2$.

With these results in hand, a typical question that might arise is - Why do we care about unitarily similar matrices or the Schur form? Yes, these matrices have desirable properties, but how are they generally used to combat specific problems? We demonstrate one answer to this question in the following example.

Example 54. Consider being given $A \in \mathbb{R}^{p\times p}$ and being asked to compute A^{1000}. Certainly, we can calculate powers of matrices, and even better, we can use computational means to arrive at the answer rather than doing anything by hand. Unfortunately, such a computation still requires a large amount of flops (floating-point operations - just additions, subtractions, multiplications, and divisions of numbers) to process. So, this may be a fairly expensive computation, even when p is not very large.

It turns out that if we can compute the Schur form of A, then this computation can be drastically reduced. Notice, for instance, that if $A = QTQ^T$, then

$$A^2 = QT\underbrace{Q^TQ}_{=\mathbb{I}}TQ^T = QTTQ^T = QT^2Q^T.$$

Continuing this, we see that

$$A^{1000} = QT^{1000}Q^T.$$

Therefore, computing powers of A can be reduced to computing powers of T. What's the difference between these two operations? Because T is upper triangular, it contains around half as many nonzero entries as A and hence, far fewer flops are needed to compute T^2 or T^{1000} than A^2 or A^{1000}. Once T^{1000} is computed, what remains is two matrix multiplications by Q and Q^T, which are quite inexpensive by comparison.

Finally, if A is unitarily similar to a diagonal matrix (i.e. unitarily diagonalizable - see the Spectral Theorem), then T is diagonal. In this case, computing powers of T is equivalent to computing the powers of the diagonal elements of T, and computing powers of a given matrix using this method is extremely quick and inexpensive.

7.5 Normal and Hermitian Operators

Definition 7.10. Let $\mathcal{V}$ be a Hilbert space with $T : \mathcal{V} \to \mathcal{V}$ bounded and linear.

1. We say that T is **normal** if T and T^* commute; that is,

$$TT^* = T^*T.$$

Said another way, T is normal if for every $v \in \mathcal{V}$, we have

$$T(T^*(v)) = T^*(T(v)).$$

2. We say that T is **Hermitian** (or **self-adjoint**) if

$$T = T^*.$$

Said another way, T is Hermitian if for every $v \in \mathcal{V}$, we have

$$T(v) = T^*(v).$$

For $A \in \mathbb{C}^{p\times p}$, these definitions become

1. A is **normal** if
$$AA^H = A^H A.$$
2. A is **Hermitian** if
$$A^H = A.$$

For $A \in \mathbb{R}^{p\times p}$, these definitions become

1. A is **normal** if
$$AA^T = A^T A.$$
2. A is **symmetric** if
$$A^T = A.$$

Comment. From these definitions, we can see that normal operators generalize many of the other classes of linear operators that we've discussed. In particular, using the definition it shouldn't be too difficult to notice that

1. Every Hermitian operator ($T^* = T$) is normal ($TT^* = T^*T$).
2. Every unitary operator ($T^*T = TT^* = \mathbb{I}$) is normal ($TT^* = T^*T$).
3. Every real, symmetric matrix ($A^T = A$) is Hermitian ($A^H = A$), and thus normal.

Example 55. Of course, not every normal matrix is Hermitian, as such a result would render the distinction between the two categories meaningless. Indeed, if we define the real 2×2 matrix

$$A = \begin{bmatrix} 0 & -2 \\ 2 & 0 \end{bmatrix}$$

so that

$$A^T = \begin{bmatrix} 0 & 2 \\ -2 & 0 \end{bmatrix}$$

and thus

$$A^T A = AA^T = 4\mathbb{I},$$

then we see that A is normal. In fact, A is neither an orthogonal (or unitary) matrix as $A^T A \neq \mathbb{I}$, nor a symmetric (or Hermitian) matrix because $A^T \neq A$. Instead, A is skew-symmetric, i.e. $A^T = -A$. More generally, every real, skew-symmetric matrix is normal because $AA^T = -A^2 = A^T A$, but not Hermitian.

Now that we have a few examples and some intuition, we will return to the general framework of normal operators (not necessarily matrices) and prove some important theorems leading up to the Spectral Theorem.

Lemma 7.9. Assume $T : \mathcal{V} \to \mathcal{V}$ is Hermitian. Then, we have $\langle T(v), v\rangle = 0$ for every $v \in \mathcal{V}$ if and only if $T = 0$.

Proof. We prove the forward direction and note that the proof of the reverse statement is similar. Assume $\langle T(v), v\rangle = 0$ for every $v \in \mathcal{V}$ where T is Hermitian. Then, using the fact that T is linear and Hermitian, we have for any choice of $u, w \in \mathcal{V}$

$$\begin{aligned}
0 &= \langle T(u+w), u+w\rangle \\
&= \langle T(u)+T(w), u+w\rangle \\
&= \cancelto{0}{\langle T(u), u\rangle} + \langle T(w), u\rangle + \langle T(u), w\rangle + \cancelto{0}{\langle T(w), w\rangle} \\
&= \langle T(w), u\rangle + \langle u, \underbrace{T^*(w)}_{T(w)}\rangle \\
&= \langle T(w), u\rangle + \langle u, T(w)\rangle \\
&= \langle T(w), u\rangle + \overline{\langle T(w), u\rangle} \\
&= 2Re(\langle T(w), u\rangle).
\end{aligned}$$

Since this is true for any $u, w \in \mathcal{V}$, we may choose $u = T(w)$ so that the result is

$$Re(\langle T(w), T(w)\rangle) = 0$$

Of course, since $\langle T(w), T(w)\rangle$ is real-valued, this just becomes

$$\|T(w)\|^2 = 0$$

for every $w \in \mathcal{V}$. Hence, $T(w) = 0$ for all $w \in \mathcal{V}$ and so $T = 0$. □

Using Lemma 7.9, we can now prove a characterization of normal operators that demonstrates that they are the only linear maps with the same norm as their adjoint for every element of their domain.

Theorem 7.10. Let $T : \mathcal{V} \to \mathcal{V}$ be a bounded linear operator. Then, T is normal if and only if

$$\|T(v)\| = \|T^*(v)\| \quad \text{for all } v \in \mathcal{V}.$$

Proof. To prove both directions simultaneously, we will establish a string of equivalences. In particular, for every $v \in \mathcal{V}$ we have

$$\begin{aligned}
\|T(v)\| = \|T^*(v)\| &\iff \|T(v)\|^2 = \|T^*(v)\|^2 \\
&\iff \langle T(v), T(v)\rangle = \langle T^*(v), T^*(v)\rangle \\
&\iff \langle T^*(T(v)), v\rangle = \langle T(T^*(v)), v\rangle \\
&\iff \langle T^*(T(v)) - T(T^*(v)), v\rangle = 0.
\end{aligned}$$

Now, define $S : \mathcal{V} \to \mathcal{V}$ by $S(v) = T^*(T(v)) - T(T^*(v))$ for every $v \in \mathcal{V}$. Then, the final equality above is simply

$$\langle S(v), v\rangle = 0$$

for every $v \in \mathcal{V}$. Furthermore, a short computation shows

$$S^* = (T^*T - TT^*)^* = T^*T - TT^* = S$$

so that S is Hermitian. Therefore, by Lemma 7.9 we find that $\langle S(v), v\rangle = 0$ for all $v \in \mathcal{V}$ is further equivalent to $S = 0$. Of course, $S = 0$ is identical to $T^*T = TT^*$, which is the definition of T being a normal operator. Thus, T is normal if and only if $\|T(v)\| = \|T^*(v)\|$, and the proof is complete. □

Theorem 7.11. Let $T : \mathcal{V} \to \mathcal{V}$ be a normal operator. Then, we have

1. If $T(v) = \lambda v$ for some $\lambda \in \mathbb{C}$ and $v \in \mathcal{V}$, then $T^*(v) = \bar{\lambda} v$. Said another way, T and T^* have complex conjugate eigenvalues corresponding to the same eigenvectors.

2. If $\lambda_1, \lambda_2 \in \mathbb{C}$ are distinct eigenvalues of T with corresponding eigenvectors v_1, v_2, then $v_1 \perp v_2$. (Note the orthogonality of eigenvectors in comparison to mere linear independence as in Lemma 7.2.)

Proof. To prove the first result, we assume $T(v) = \lambda v$ so that $\|T(v) - \lambda v\|^2 = 0$. Then, using Theorem 7.10 the normality of T implies $\|T^*(v)\| = \|T(v)\|$. Using this and the adjoint property, we find

$$\begin{aligned}
\|T(v) - \lambda v\|^2 &= \langle T(v) - \lambda v, T(v) - \lambda v\rangle \\
&= \|T(v)\|^2 - \bar{\lambda}\langle v, T(v)\rangle - \lambda\langle T(v), v\rangle + \lambda^2\langle v, v\rangle \\
&= \|T^*(v)\|^2 - \bar{\lambda}\langle T^*(v), v\rangle - \lambda\langle v, T^*(v)\rangle + \lambda^2\langle v, v\rangle \\
&= \|T^*(v)\|^2 - \langle T^*(v), \bar{\lambda} v\rangle - \langle \bar{\lambda} v, T^*(v)\rangle + \lambda^2\langle v, v\rangle \\
&= \langle T^*(v) - \bar{\lambda} v, T^*(v) - \bar{\lambda} v\rangle \\
&= \|T^*(v) - \bar{\lambda} v\|^2.
\end{aligned}$$

As this quantity is zero, we find $T^*(v) = \bar{\lambda} v$ and the proof of the first result is complete.

Now, to prove the second result we assume there are $\lambda_1, \lambda_2 \in \mathbb{C}$ and $v_1, v_2 \in \mathcal{V}$ such that

$$T(v_1) = \lambda_1 v_1 \qquad \text{and} \qquad T(v_2) = \lambda_2 v_2.$$

Then, the first result implies $T^*(v_1) = \bar{\lambda}_1 v_1$. Therefore, we use the adjoint property to arrive at

$$\lambda_2\langle v_1, v_2\rangle = \langle v_1, \lambda_2 v_2\rangle = \langle v_1, T(v_2)\rangle = \langle T^*(v_1), v_2\rangle = \langle \bar{\lambda}_1 v_1, v_2\rangle = \lambda_1\langle v_1, v_2\rangle.$$

Subtracting, this becomes

$$(\lambda_2 - \lambda_1)\langle v_1, v_2\rangle = 0.$$

Finally, if $\lambda_1 \neq \lambda_2$, then $\langle v_1, v_2\rangle = 0$ follows immediately. □

Because they are identical to their own adjoints, Hermitian operators also possess special eigenvalue properties, which will be useful later on.

Theorem 7.12. Let $T : \mathcal{V} \to \mathcal{V}$ be Hermitian. Then, all eigenvalues of T are real.

Proof. Assume $T(v) = \lambda v$ for some $\lambda \in \mathbb{C}$ and $v \in \mathcal{V} \setminus \{0\}$. Then, we find

$$\langle T(v), v\rangle = \langle \lambda v, v\rangle = \bar{\lambda}\langle v, v\rangle = \bar{\lambda}\|v\|^2.$$

Similarly, because T is Hermitian, and thus $T^* = T$, we also find

$$\langle v, T^*(v)\rangle = \langle v, T(v)\rangle = \langle v, \lambda v\rangle = \lambda\langle v, v\rangle = \lambda\|v\|^2.$$

However, these two expressions must be equal since combining these equalities with the adjoint property guarantees

$$\bar{\lambda}\|v\|^2 = \langle T(v), v\rangle = \langle v, T^*(v)\rangle = \lambda\|v\|^2.$$

Since $v \neq 0$, we divide this by the norm of v to find $\bar{\lambda} = \lambda$, which implies that λ is real. Since λ is an arbitrary eigenvalue, all eigenvalues must be real. □

7.6 Cholesky decomposition

In addition to possessing real eigenvalues, certain types of Hermitian matrices also give rise to a special matrix factorization that arises frequently within applications, such as the study of covariance (or correlation) matrices in statistics. Before stating the theorem representing this result, a new definition is needed.

Definition 7.11. We say a Hermitian operator $T : \mathcal{V} \to \mathcal{V}$ is **positive definite** if

$$\langle v, T(v)\rangle > 0$$

for every $v \in \mathcal{V} \setminus \{0\}$. Analogously, a Hermitian matrix $A \in \mathbb{C}^{p\times p}$ is **positive definite** if

$$x^H A x > 0$$

for every $x \in \mathbb{C}^p \setminus \{0\}$, while a symmetric matrix $B \in \mathbb{R}^{p\times p}$ is **positive definite** if

$$x^T B x > 0$$

for every $x \in \mathbb{R}^p \setminus \{0\}$, respectively.

Though this property is difficult to check since it must be verified for all such vectors, it is equivalent to another condition that is easier to determine. The proof of this equivalence utilizes the celebrated **Rayleigh quotient**, namely

$$R_T(v) = \frac{\langle v, T(v)\rangle}{\langle v, v\rangle} \quad \text{or} \quad R_A(x) = \frac{x^H A x}{\|x\|_2^2}.$$

Theorem 7.13. A Hermitian matrix $A \in \mathbb{C}^{p\times p}$ is positive definite if and only if all eigenvalues of A are positive.

Proof. We begin by letting $\lambda \in \mathbb{C}$ and $x \in \mathbb{C}^p \setminus \{0\}$ be any eigenpair of A so that

$$Ax = \lambda x.$$

Multiplying by x^H on the left yields

$$x^H Ax = x^H \lambda x = \lambda \|x\|_2^2,$$

which, upon dividing by $\|x\|_2^2$, becomes

$$\lambda = \frac{x^H Ax}{\|x\|_2^2}.$$

Therefore, we see that $\lambda > 0$ if and only if $x^H Ax > 0$. Thus, if A is positive definite then $x^H Ax > 0$ for every eigenvector, and every eigenvalue is positive.

Additionally, because A is Hermitian, the Spectral decomposition (Theorem 7.19), which we will justify in the next section, implies that A is unitarily diagonalizable. Hence, there is an orthonormal basis of $\mathbb{C}^p$ consisting only of eigenvectors of A. Thus, every $x \in \mathbb{C}^p$ can be written as a linear combination of eigenvectors of A, and there is an orthonormal set $\{v_1, ..., v_p\} \subseteq \mathbb{C}^p$ such that for any $x \in \mathbb{C}^p$, we have unique $\alpha_j \in \mathbb{C}$ for every $j = 1, ..., p$ such that

$$x = \sum_{j=1}^{p} \alpha_j v_j.$$

Defining the Kronecker delta function

$$\delta_{ij} = \begin{cases} 0 & i \neq j \\ 1 & i = j \end{cases},$$

the orthonormal nature of $\{v_1, ..., v_p\}$ can be expressed as

$$v_i^H v_j = \delta_{ij}$$

for every $i, j = 1, ..., p$. Using the above basis expansion in terms of eigenvectors, we find

$$\begin{aligned} x^H Ax &= \left(\sum_{j=1}^{p} \alpha_j v_j \right)^H A \left(\sum_{k=1}^{p} \alpha_k v_k \right) \\ &= \sum_{j=1}^{p} \sum_{k=1}^{p} \bar{\alpha}_j \alpha_k v_j^H A v_k \\ &= \sum_{j=1}^{p} \sum_{k=1}^{p} \bar{\alpha}_j \alpha_k \lambda_k v_j^H v_k \\ &= \sum_{j=1}^{p} \sum_{k=1}^{p} \bar{\alpha}_j \alpha_k \lambda_k \delta_{jk} \\ &= \sum_{j=1}^{p} |\alpha_j|^2 \lambda_j. \end{aligned}$$

If $x \neq 0$, then $\alpha_\ell > 0$ for some $\ell = 1, ..., p$. Therefore, if every eigenvalue is positive, we find that $x^H Ax > 0$ for every $x \neq 0$. □

Another property of positive definite matrices is that they admit a specific matrix decomposition similar to the LU decomposition from Linear Algebra, but requiring knowledge of a single triangular matrix only.

Theorem 7.14 (Cholesky)**.** Assume $A \in \mathbb{C}^{p\times p}$ is (Hermitian) positive definite. Then, there is a unique lower triangular matrix $L \in \mathbb{C}^{p\times p}$ such that

$$A = LL^H.$$

This representation is referred to as the **Cholesky decomposition**.

Instead of proving this theorem, we will prove its real counterpart, namely:

Corollary 7.3. Assume $A \in \mathbb{R}^{p\times p}$ is (symmetric) positive definite. Then, there is a unique lower triangular matrix $L \in \mathbb{R}^{p\times p}$ such that

$$A = LL^T.$$

This representation is also referred to as the **Cholesky decomposition**.

Proof. First, we assume that A is positive definite. In view of Theorem 7.13, A has only positive eigenvalues, and hence is invertible. By a generalization of the LU Factorization (known as the LDU Factorization) [5], there are unique diagonal $D \in \mathbb{R}^{p\times p}$, upper triangular $U \in \mathbb{R}^{p\times p}$, and lower triangular $K \in \mathbb{R}^{p\times p}$ with $K_{jj} = U_{jj} = 1$ for every $j = 1, ..., p$ such that

$$A = KDU.$$

Computing A^T we find

$$A^T = U^T D^T K^T = U^T D K^T$$

and using the symmetry of A, this implies

$$U^T D K^T = A^T = A = KDU.$$

Since U^T is lower triangular and K^T is upper triangular, we have actually found two LDU decompositions, but since such a decomposition is unique, the matrices in each must be equal. Hence, we find $U = K^T$ and therefore

$$A = KDK^T.$$

Notice further that D must have positive diagonal entries. Indeed, since A is positive definite, $x^T Ax > 0$ for all $x \neq 0$. Thus, $x^T KDK^T x > 0$ or

$$(K^T x)^T D(K^T x) > 0$$

for all $x \neq 0$. Since K is lower triangular with non-zero diagonal entries, it is invertible and a simple change of variables $y = K^T x$ shows that

$$y^T Dy > 0$$

for all $y \neq 0$. Hence, D is also positive definite and has only positive eigenvalues. Since D is diagonal, these are exactly the diagonal entries, which must then be positive.

With this, we merely construct $B = D^{1/2}$, which is the diagonal matrix defined by $b_{ii} = \sqrt{d_{ii}}$, so that $B^2 = D$. Thus, we have the representation

$$A = KB^2K^T$$

where K is lower triangular and B is diagonal. Finally, define $L = KB$ so that L is lower triangular and $L^T = (KB)^T = BK^T$. Then, we have

$$A = (KB)(BK^T) = LL^T.$$

□

Comment. These theorems can actually be made into equivalences (i.e., iff statements), though we have only proved one direction.

7.7 Spectral Theorem

Though Theorem 7.10 provided an equivalent condition to the normality property of a linear operator, it is certainly not the only characterization, especially for normal matrices. In fact, one of the most celebrated theorems in Linear Algebra (namely, the Spectral Theorem for real, symmetric matrices - see Theorem 13 in the review section) has a generalization that states a well-known equivalence of normal operators of a certain type defined on a Hilbert space, which we will present shortly. First, though, we must define some specific terminology.

Definition 7.12. Let $\mathcal{V}$ and $\mathcal{W}$ be Banach spaces and $T : \mathcal{V} \to \mathcal{W}$ be a linear operator. We say T is **compact** if for every bounded sequence $\{v_n\}_{n=1}^{\infty} \subset \mathcal{V}$, the sequence $\{T(v_n)\}_{n=1}^{\infty} \subset \mathcal{W}$ has a convergent subsequence.

The compactness property of a linear operator defined on a Banach space essentially indicates that it behaves like a matrix. In fact, if the spaces $\mathcal{V}$ and $\mathcal{W}$ in the definition above are finite-dimensional, then any $T : \mathcal{V} \to \mathcal{W}$ is necessarily compact. Because of this, many results from Linear Algebra can often be extended to compact operators using similar arguments. In particular, as the next result shows, every such operator must be bounded. Contrastingly, the study of non-compact linear operators on infinite-dimensional spaces often necessitates very different approaches and can give rise to non-intuitive results.

Example 56. We will demonstrate the basic notion of a compact operator using a few examples.

1. Consider the identity operator $I : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$. Then, I is not compact. To show this, we consider the bounded sequence $e_n \in \ell^2$ for every $n \in \mathbb{N}$ where

$$e_n = (0, 0, 0, ..., 1, ..., 0, 0, ...)$$

is the sequence with a 1 in the nth term and zeros in all other terms. Thus, $\{e_n\}_{n=1}^{\infty} = \{e_1, e_2, ...\}$ is a bounded sequence of elements of ℓ^2, each of which is itself a square-summable sequence of real numbers. However, we note that

$$\|I(e_n) - I(e_m)\| = \|e_n - e_m\| = \sqrt{2}$$

for $n \neq m$. Thus, no subsequence of $I(e_n)$ can be Cauchy, and hence, no subsequence can converge. Therefore, we have created a bounded sequence $\{e_n\}_{n=1}^{\infty}$ such that the sequence $\{I(e_n)\}_{n=1}^{\infty}$ does not have a convergent subsequence in ℓ^2, which implies that I is not compact.

2. Fix a value of $N \in \mathbb{N}$ and consider the associated linear operator $T_N : \ell^2(\mathbb{R}) \to \ell^2(\mathbb{R})$ defined by

$$T_N(x) = (x_1, ..., x_N, 0, 0, ...).$$

Then, T_N is compact. Indeed, given any bounded sequence of elements $x^{(n)} \in \ell^2$ for $n \in \mathbb{N}$, we see that $T_N(x^{(n)}) \in \mathbb{R}^N$ is actually a sequence of real vectors for every $n \in \mathbb{N}$. Additionally, $T_N(x^{(n)})$ satisfies

$$\left|T_N(x^{(n)})\right|^2 = |x_1^{(n)}|^2 + \cdots + |x_N^{(n)}|^2 = \sum_{k=1}^{N} |x_k^{(n)}|^2 \leq \sum_{k=1}^{\infty} |x_k^{(n)}|^2 = \|x^{(n)}\|_{\ell^2}^2,$$

which is bounded. Therefore, $T_N(x^{(n)})$ is a bounded sequence of vectors in $\mathbb{R}^N$. By the Bolzano-Weierstrauss Theorem (Appendix: Theorem 9.1), every bounded sequence of real vectors has a convergent subsequence. Hence, $T_N(x^{(n)})$ has a convergent subsequence. This, then implies that T is compact.

3. Let $K \in C([0,1] \times [0,1])$ and $T : L^2([0,1]) \to L^2([0,1])$ be defined by

$$T(f) = \int_0^1 K(x,y) f(y) dy.$$

Then, the integral operator T is compact. We will not prove this, but the importance of such a result will be readily apparent to anyone with a familiarity with differential or integral equations.

Next, we show that all compact operators are necessarily bounded.

Lemma 7.15. Let $\mathcal{V}$ be a Hilbert space and $T : \mathcal{V} \to \mathcal{V}$ be compact. Then, T is bounded. Consequently, T^* exists.

Proof. We prove the contrapositive. Assume T is not bounded, so that $\|T\| = \infty$. Then, there is a sequence $v_n \in \mathcal{V}$ with $\|v_n\| \leq 1$ for every $n \in \mathbb{N}$ and $\|T(v_n)\| \to \infty$ as $n \to \infty$. Then, the sequence $\{T(v_n)\}_{n=1}^{\infty}$ cannot possess a convergent subsequence as each term grows without bound. Hence, this implies that T is not compact, and proves the first conclusion. The second conclusion merely follows by the boundedness of T. □

Definition 7.13. Let $\mathcal{V}$ and $\mathcal{W}$ be vector spaces and $T : \mathcal{V} \to \mathcal{W}$ be a linear operator. For any subspace $\mathcal{V}_0 \subseteq \mathcal{V}$, we define the **restriction** of T to the domain $\mathcal{V}_0$ by the linear operator $T_0 : \mathcal{V}_0 \to \mathcal{W}$ satisfying

$$T_0(v) = T(v)$$

for every $v \in \mathcal{V}_0$. Instead of writing this operator as T_0, the restriction of T to the domain $\mathcal{V}_0$ is sometimes written as $T|_{\mathcal{V}_0}$ to explicitly represent the domain.

It should be noted that the restriction T_0 of an operator T generally maintains each of the properties of the original operator. In particular, the norm of a restriction operator is always bounded above by the norm of the original operator so that $\|T_0\| \leq \|T\|$. Also, if T is compact, then so is T_0. Finally, if $T : \mathcal{V} \to \mathcal{V}$ is normal (unitary, Hermitian) and the associated adjoint operator is defined on a subspace $\mathcal{V}_0$ with $T(v), T^*(v) \in \mathcal{V}_0$ for every $v \in \mathcal{V}_0$, then $T_0 : \mathcal{V}_0 \to \mathcal{V}_0$ is also normal (unitary, Hermitian). The proofs of these assertions are left as exercises (though they can be justified without much effort), and they will be particularly useful properties in proving the Spectral Theorem.

Before stating the main result, we need one additional lemma concerning the spectral radius of normal operators.

Lemma 7.16. Let $\mathcal{V}$ be a Hilbert space and $T : \mathcal{V} \to \mathcal{V}$ be a normal, bounded linear operator. Then,

$$\rho(T) = \|T\|.$$

In particular, there exists an eigenvalue of T, denoted λ, such that

$$|\lambda| = \|T\|.$$

Proof. This proof requires an explicit formula for the spectral radius, which we state here and direct the reader to [9] for more information. In particular, for a bounded linear operator T, we can write

$$\rho(T) = \lim_{n\to\infty} \|T^n\|^{1/n}. \tag{7.2}$$

Thus, we will now focus on deriving a formula for T^n and using (7.2).

Let $S : \mathcal{V} \to \mathcal{V}$ be defined by $S = T^*T$ and note that (due to Problem 5.8)

$$S^* = (T^*T)^* = T^*T = S$$

so that S is Hermitian. With this property, we use Cauchy-Schwarz and Theorem 5.16(d) to compute for any $v \in \mathcal{V}$

$$\|S(v)\|^2 = \langle S(v), S(v)\rangle = \langle S^*(S(v)), v\rangle = \langle S^2(v), v\rangle \leq \|S^2(v)\| \cdot \|v\| \leq \|S^2\| \cdot \|v\|^2.$$

Assuming $v \neq 0$, dividing by $\|v\|^2$, and taking the supremum over all such $v \in \mathcal{V}$, we find $\|S\|^2 \leq \|S^2\|$. Additionally, for any $v \in \mathcal{V}$ we use Theorem 5.16(d) to find

$$\|S^2(v)\| = \|S(S(v))\| \leq \|S\| \cdot \|S(v)\| \leq \|S\|^2\|v\|,$$

which then implies $\|S^2\| \leq \|S\|^2$. Putting these together, we conclude

$$\|S^2\| = \|S\|^2.$$

By induction, we can continue this to find $\|S^{2^m}\| = \|S\|^{2^m}$ or

$$\left\|(T^*T)^{2^m}\right\| = \|T^*T\|^{2^m} \tag{7.3}$$

for any $m \in \mathbb{N}$.

Next, we show $\|S\| = \|T\|^2$. Indeed, using Problem 5.9 we find

$$\|S\| = \|T^*T\| \leq \|T^*\| \cdot \|T\| = \|T\|^2.$$

On the other hand, we have for any $v \in \mathcal{V}$

$$\|T(v)\|^2 = \langle T(v), T(v)\rangle = \langle T^*(T(v)), v\rangle \leq \|T^*T\| \cdot \|v\|^2 = \|S\| \cdot \|v\|^2,$$

which implies $\|T\| \leq \sqrt{\|S\|}$ or $\|T\|^2 \leq \|S\|$. Hence, we have $\|S\| = \|T\|^2$ or

$$\|T^*T\| = \|T\|^2. \tag{7.4}$$

Finally, we use the normality of T to find

$$\left(T^{2^m}\right)^* T^{2^m} = (T\cdots T)^*(T\cdots T) = (T^*\cdots T^*)(T\cdots T) = (T^*T\cdots T^*T) = (T^*T)^{2^m}$$

Therefore, using this equality along with (7.3), (7.4), and Problem 7.11, we have

$$\left\|T^{2^m}\right\|^2 = \left\|\left(T^{2^m}\right)^* T^{2^m}\right\| = \left\|(T^*T)^{2^m}\right\| = \|T^*T\|^{2^m} = \|T\|^{2^{m+1}}.$$

Thus, $\left\|T^{2^m}\right\| = \|T\|^{2^m}$, and we can now compute the spectral radius by taking the limit along $n = 2^m$ as $m \to \infty$. This yields

$$\rho(T) = \lim_{m\to\infty} \left\|T^{2^m}\right\|^{1/2^m} = \lim_{m\to\infty} \|T\|^{2^m/2^m} = \|T\|.$$

Finally, Theorem 7.1 guarantees that there exists an eigenvalue $\lambda \in \mathbb{C}$ satisfying $|\lambda| = \rho(T)$. Thus, $|\lambda| = \|T\|$ and the proof is complete. □

With this result established, we can now prove the Spectral Theorem.

Theorem 7.17 (Spectral Theorem)**.** Let $\mathcal{V}$ be a Hilbert space and $T : \mathcal{V} \to \mathcal{V}$ be a linear operator. Then, T is compact and normal if and only if there exists an orthonormal sequence of eigenvectors $e_n \in \mathcal{V}$ with corresponding eigenvalues $\lambda_n \in \mathbb{C}$ such that

$$T(v) = \sum_{n=1}^{\infty} \lambda_n \langle e_n, v\rangle e_n \tag{7.5}$$

for every $v \in \mathcal{V}$. Furthermore, if λ_n has infinitely many terms, then $\lambda_n \to 0$ as $n \to \infty$.

Proof. We first prove the forward implication. Notice that if $T = 0$, then clearly $\lambda = 0$ is the only eigenvalue so that any $v \in \mathcal{V} \setminus \{0\}$ serves as a corresponding eigenvector, and (7.5) holds. Thus, suppose $T \neq 0$. Then, by Lemma 7.16 there exists an eigenvalue $\lambda_1 \in \mathbb{C}$ such that $|\lambda_1| = \|T\| > 0$ with corresponding eigenvector $e_1 \in \mathcal{V} \setminus \{0\}$ satisfying $\|e_1\| = 1$. Now, define

$$\mathcal{V}_1 = \operatorname{span}\{e_1\}^{\perp}.$$

If $v \in \mathcal{V}_1$, then $\langle v, e_1 \rangle = 0$ and by Theorem 7.11

$$\langle T(v), e_1 \rangle = \langle v, T^*(e_1) \rangle = \langle v, \bar{\lambda}_1 e_1 \rangle = \bar{\lambda}_1 \langle v, e_1 \rangle = 0$$

and

$$\langle e_1, T^*(v) \rangle = \langle T(e_1), v \rangle = \langle \lambda_1 e_1, v \rangle = \bar{\lambda}_1 \langle e_1, v \rangle = 0.$$

Thus, $T(v), T^*(v) \in \mathcal{V}_1$. Let $T_1 = T|_{\mathcal{V}_1}$, which is also a normal, compact operator with $\|T_1\| \leq \|T\| = |\lambda_1|$. Using T_1 and Lemma 7.16, we may repeat this procedure to construct $\lambda_2 = \|T_1\|$, e_2 with $\|e_2\| = 1$, $\mathcal{V}_2 = \operatorname{span}\{e_1, e_2\}^{\perp}$, and T_2, then continue to obtain a sequence of eigenvalues $\lambda_n \in \mathbb{C}$, unit eigenvectors $e_n \in \mathcal{V}$, subspaces $\mathcal{V}_n \subseteq \mathcal{V}$, and restricted normal, compact operators $T_n : \mathcal{V}_n \to \mathcal{V}_n$ with $\|T_n\| \leq \|T_{n-1}\| = |\lambda_n|$ for $n \in \mathbb{N}$. For any $v \in \mathcal{V}$ and $N \in \mathbb{N}$, we define the remainder of an N-term approximation of v by

$$u_N = v - \sum_{n=1}^{N} \langle e_n, v \rangle e_n$$

so that $u_N \in \mathcal{V}_N$ and rearranging

$$v = \sum_{n=1}^{N} \langle e_n, v \rangle e_n + u_N. \tag{7.6}$$

If $T_N = 0$ for some $N \in \mathbb{N}$, then the operator sequence terminates and (7.5) holds because $T_N(u_N) = 0$ in the final calculation below. Otherwise, we obtain a sequence of eigenvalues $\{\lambda_n\}_{n=1}^{\infty}$ and claim that $\lambda_n \to 0$ as $n \to \infty$. Indeed, if $\lambda_n \not\to 0$ as $n \to \infty$, then there exists $\epsilon > 0$ and $M \in \mathbb{N}$ such that $|\lambda_n| > \epsilon$ for $n > M$. Then, it follows by orthonormality of the eigenvectors and the Pythagorean Theorem (cf. Problem 4.16) that for $n, m > M$ and $n \neq m$

$$\|T(e_n) - T(e_m)\|^2 = \|\lambda_n e_n - \lambda_m e_m\|^2 = |\lambda_n|^2 + |\lambda_m|^2 > \epsilon^2.$$

This then shows that $T(e_n)$ has no convergent subsequence, and since e_n is a bounded sequence in $\mathcal{V}$, it further contradicts the assumption that T is compact. Hence, we find $\lambda_n \to 0$ as $n \to \infty$.

Now, because each term in (7.6) is orthogonal, their inner products vanish and thus by the Pythagorean Theorem (cf. Problem 4.16)

$$\|v\|^2 = \sum_{n=1}^{N} |\langle e_n, v \rangle|^2 + \|u_N\|^2.$$

This implies $\|u_N\| \leq \|v\|$, and thus

$$\|T(u_N)\| = \|T_N(u_N)\| \leq \|T_N\| \|u_N\| \leq |\lambda_N| \|v\| \to 0$$

as $N \to \infty$, which means $T(u_N) \to 0$ in $\mathcal{V}$ as $N \to \infty$. Using (7.6) and the linearity of T, we find

$$T(v) = T\left(\sum_{n=1}^{N} \langle e_n, v\rangle e_n + u_N\right) = \sum_{n=1}^{N} \langle e_n, v\rangle T(e_n) + T(u_N)$$

for any $v \in \mathcal{V}$ and $N \in \mathbb{N}$. Since $T(u_N)$ tends to zero, we take the limit as $N \to \infty$ in the right side and conclude

$$T(v) = \sum_{n=1}^{\infty} \langle e_n, v\rangle T(e_n) = \sum_{n=1}^{\infty} \lambda_n \langle e_n, v\rangle e_n.$$

To establish the reverse implication, we will only prove that T defined by (7.5) is normal. The compactness of this operator can be shown independently. We consider $v, w \in \mathcal{V}$ and use Theorem 4.17 to write

$$\begin{aligned}
\langle T(v), w\rangle &= \left\langle \sum_{n=1}^{\infty} \lambda_n \langle e_n, v\rangle e_n, w \right\rangle \\
&= \sum_{n=1}^{\infty} \bar{\lambda}_n \overline{\langle e_n, v\rangle} \langle e_n, w\rangle \\
&= \sum_{n=1}^{\infty} \bar{\lambda}_n \langle v, e_n\rangle \langle e_n, w\rangle \\
&= \left\langle v, \sum_{n=1}^{\infty} \bar{\lambda}_n \langle e_n, w\rangle e_n \right\rangle.
\end{aligned}$$

Thus, we find the adjoint of T is

$$T^*(w) = \sum_{n=1}^{\infty} \bar{\lambda}_n \langle e_n, w\rangle e_n.$$

Finally, we compute the composition of T^* and T and use the orthogonality of e_k and e_n for $k \neq n$ to find for any $v \in \mathcal{V}$

$$\begin{aligned}
T^*(T(v)) &= \sum_{n=1}^{\infty} \bar{\lambda}_n \langle e_n, T(v)\rangle e_n \\
&= \sum_{n=1}^{\infty} \bar{\lambda}_n \left\langle e_n, \sum_{k=1}^{\infty} \lambda_k \langle e_k, v\rangle e_k \right\rangle e_n \\
&= \sum_{n=1}^{\infty} \sum_{k=1}^{\infty} \bar{\lambda}_n \lambda_k \langle e_k, v\rangle \langle e_n, e_k\rangle e_n \\
&= \sum_{n=1}^{\infty} \bar{\lambda}_n \lambda_n \langle e_n, v\rangle e_n \\
&= \sum_{n=1}^{\infty} |\lambda_n|^2 \langle e_n, v\rangle e_n
\end{aligned}$$

and the same computation holds for $T(T^*(v)) = T^*(T(v))$ which yields the normality of T.

□

Theorem 7.18. Let $\mathcal{V}$ be a Hilbert space with an orthonormal basis and $T : \mathcal{V} \to \mathcal{V}$ be a linear operator. Then, T compact and normal if and only if there exists an orthonormal basis for $\mathcal{V}$ consisting only of eigenvectors d_n of T such that

$$T(v) = \sum_{n=1}^{\infty} \lambda_n \langle d_n, v \rangle d_n$$

for every $v \in \mathcal{V}$, where $\lambda_n \in \mathbb{C}$ are eigenvalues of T including zeros.

Proof. We will assume that we're dealing with an infinite dimensional Hilbert space with the proof of the finite dimensional case following analogously merely by replacing infinite sets and series with finite ones. Let $\{e_n\}_{n=1}^{\infty} \subset \mathcal{V}$ be the orthonormal sequence of eigenvectors guaranteed by Theorem 7.17. We first note that $e_n \in \text{Ker}(T)$ if and only if $\lambda_n = 0$. Certainly, if $e_n \in \text{Ker}(T)$, then $\lambda_n e_n = T(e_n) = 0$ and since $e_n \neq 0$, it follows that $\lambda_n = 0$. Additionally, $\lambda_n = 0$ implies $e_n \in \text{Ker}(T)$ because $T(e_n) = \lambda_n e_n = 0$.

Next, we show that an element $v \in \mathcal{V}$ is orthogonal to every $e_n \notin \text{Ker}(T)$ if and only if $v \in \text{Ker}(T)$. Indeed, if $\langle e_n, v \rangle = 0$ for every $e_n \notin \text{Ker}(T)$, then by (7.5) we see that $T(v) = 0$ because the remaining terms in the sum correspond to zero eigenvalues. Alternatively, if $T(v) = 0$ then by the orthogonality of e_n and the Pythagorean Theorem (cf. 4.16)

$$0 = \|T(v)\|^2 = \sum_{n=1}^{\infty} |\lambda_n|^2 \cdot |\langle e_n, v \rangle|^2.$$

The nonnegativity of the terms in this sum implies that each must be exactly zero. As we saw above, $\lambda_n = 0$ implies $e_n \in \text{Ker}(T)$ because $T(e_n) = \lambda_n e_n = 0$. Hence, for all $e_n \notin \text{Ker}(T)$, we find $\langle e_n, v \rangle = 0$.

Now, since $\mathcal{V}$ has an orthonormal basis, we can construct an orthonormal basis for the subspace $\text{Ker}(T) \subseteq \mathcal{V}$. Call this basis $G = \{g_n : n \in \mathbb{N}\}$. Then, because $T(g_n) = 0$, we see that g_n is an eigenvector of T with corresponding eigenvalue $\lambda_n = 0$. Additionally, by the orthogonality property shown above, it follows that $g_n \perp e_m$ for all $n, m \in \mathbb{N}$. We then define

$$D = E \cup G$$

where $E = \{e_n : n \in \mathbb{N}\}$ and label the elements of D as d_n for $n \in \mathbb{N}$. Notice that each d_n is an eigenvector of T and for any $v \in \mathcal{V}$, we have

$$T(v) = \sum_{n=1}^{\infty} \lambda_n \langle e_n, v \rangle e_n = \sum_{n=1}^{\infty} \lambda_n \langle d_n, v \rangle d_n.$$

Finally, because D is an orthonormal set, we merely need to show that it is a Schauder basis for $\mathcal{V}$ to complete the proof. Of course, the linear independence property follows from the orthogonality of the set, so we focus on the spanning property. Let $v \in \mathcal{V}$ be given. Then, we compute

$$T(v) = \sum_{n=1}^{\infty} \lambda_n \langle e_n, v \rangle e_n = \sum_{n=1}^{\infty} \langle e_n, v \rangle T(e_n)$$

so that by linearity and continuity (i.e. boundedness) of T, we find

$$T\left(v - \sum_{n=1}^{\infty} \langle e_n, v\rangle e_n\right) = 0.$$

Therefore, $v - \sum_{n=1}^{\infty}\langle e_n, v\rangle e_n \in \text{Ker}(T)$ and can be expressed as the linear combination of elements of G. This means that there are $\alpha_n \in \mathbb{C}$ such that

$$v = \sum_{n=1}^{\infty} \left[\langle e_n, v\rangle e_n + \alpha_n g_n\right]$$

which is merely a linear combination of elements of D. Thus, D is a Schauder basis for $\mathcal{V}$ and the proof is complete. □

Comment. As an application of this theorem, consider solving the following problem. Let $\mathcal{V}$ be a Hilbert space with an orthonormal basis. Given a compact, normal operator $T : \mathcal{V} \to \mathcal{V}$ and $w \in \mathcal{V}$, we wish to find $u \in \mathcal{V}$ such that $T(u) = w$. This problem is ubiquitous throughout the mathematical sciences, as it represents solving

1. $Ax = b$ for a symmetric matrix in Linear Algebra,
2. $u'' = f$ in the study of ODEs,
3. $\Delta u = f$ in the study of PDEs, and
4. $\int_0^1 K(x, y)u(y)\, dy = f(x)$ for a symmetric kernel K in the study of integral equations.

However, each of these problems can be solved in this general context of $T(u) = w$ using the same unified framework. In particular, since T is compact and normal, we know from the Spectral Theorem (more specifically, Theorem 7.18) that

$$T(u) = \sum_{n=1}^{\infty} \lambda_n \langle d_n, u\rangle d_n$$

where we can determine the orthonormal eigenpairs (λ_n, d_n) from the given operator. Now, since $\{d_n\}_{n=1}^{\infty}$ is an orthonormal basis for $\mathcal{V}$ we can decompose w as

$$w = \sum_{n=1}^{\infty} \langle d_n, w\rangle d_n.$$

Here, the coefficients $\langle d_n, w\rangle$ are merely the coordinates of w with respect to this basis. Thus, the problem $T(u) = w$ merely becomes

$$\sum_{n=1}^{\infty} \lambda_n \langle d_n, u\rangle d_n = \sum_{n=1}^{\infty} \langle d_n, w\rangle d_n.$$

In general, solving an equation involving two infinite sums is problematic, but because the d_n vectors are orthonormal it's significantly simpler. In particular,

taking the inner product of both sides of the equation with d_k, for a fixed k and using the orthonormality property, the equation becomes

$$\sum_{n=1}^{\infty} \lambda_n \langle d_n, u\rangle \langle d_k, d_n\rangle = \sum_{n=1}^{\infty} \langle d_n, w\rangle \langle d_k, d_n\rangle$$

and thus

$$\sum_{n=1}^{\infty} \lambda_n \langle d_n, u\rangle \delta_{kn} = \sum_{n=1}^{\infty} \langle d_n, w\rangle \delta_{kn}.$$

Due to the Kronecker delta, each term of the sum vanishes except for the kth term, and we find

$$\lambda_k \langle d_k, u\rangle = \langle d_k, w\rangle.$$

Therefore, if $\lambda_k \neq 0$, we conclude

$$\langle d_k, u\rangle = \frac{\langle d_k, w\rangle}{\lambda_k}.$$

Finally, we can decompose u using the same d_n basis, and with this representation for the coordinates we ultimately have an exact expression for the solution

$$u = \sum_{n=1}^{\infty} \langle d_n, u\rangle d_n = \sum_{n=1}^{\infty} \frac{\langle d_n, w\rangle}{\lambda_n} d_n,$$

which is written in terms of λ_n, d_n and w, each of which we know.

This idea is, in fact, the basis for a variety of remarkable numerical methods for solving PDEs, called (unsurprisingly) Spectral methods. The extremely useful characteristic of these numerical methods is that, unlike Finite Difference or Finite Element methods, they can display *exponential* (rather than polynomial) convergence properties for certain classes of problems.

Next, we specialize the Spectral Theorem for complex matrices.

Theorem 7.19 (Spectral Theorem for $\mathbb{C}^{p\times p}$)**.** Let $A \in \mathbb{C}^{p\times p}$ be given. Then, A is normal if and only if A is unitarily diagonalizable, meaning there are unitary $U \in \mathbb{C}^{p\times p}$ and diagonal $D \in \mathbb{C}^{p\times p}$ such that $A = UDU^H$.

Proof. Because $\mathbb{C}^p$ is a Hilbert space with inner product

$$\langle x, y\rangle = x^H y = \sum_{n=1}^{p} \bar{x}_k y_k$$

and orthonormal basis

$$B = \{e_k : k = 1, ..., p\},$$

and $T : \mathbb{C}^p \to \mathbb{C}^p$ defined by

$$T(x) = Ax$$

for all $x \in \mathbb{C}^p$ (where A is the given matrix) is a compact normal operator, this result follows as an immediate corollary of the Spectral Theorem (more specifically, Theorem 7.18). □

To better explain the connection between Theorems 7.18 and 7.19, we first need to define a bit of notation. Given $u, v \in \mathbb{C}^p$, define the **outer product** of these vectors by the $p \times p$ matrix

$$u \otimes v = uv^H.$$

Alternatively, we can define this matrix entrywise by

$$(u \otimes v)_{ij} = u_i \bar{v}_j$$

or in the standard visual representation of matrix multiplication as

$$u \otimes v = \begin{bmatrix} u_1 \\ \vdots \\ u_p \end{bmatrix} \begin{bmatrix} \bar{v}_1 & \dots & \bar{v}_p \end{bmatrix} = \begin{bmatrix} u_1\bar{v}_1 & \dots & u_1\bar{v}_p \\ \vdots & \ddots & \vdots \\ u_p\bar{v}_1 & \dots & u_p\bar{v}_p \end{bmatrix}.$$

Then, we can use this notation to rewrite the spectral decomposition of any normal $A \in \mathbb{C}^{p\times p}$.

Indeed, if $A \in \mathbb{C}^{p\times p}$ satisfies $A = UDU^H$. Since we know that these matrices arise from eigenvalues and eigenvectors of A, we let

$$U = [v_1, \dots, v_p] \qquad \text{and} \qquad D = \operatorname{diag}(\lambda_1, \dots, \lambda_p)$$

where $Av_j = \lambda_j v_j$ for all $j = 1, \dots, p$. Then, we compute

$$DU^H = \begin{bmatrix} \lambda_1 & 0 & \dots \\ 0 & \ddots & 0 \\ 0 & \dots & \lambda_p \end{bmatrix} \begin{bmatrix} v_1^H \\ \vdots \\ v_p^H \end{bmatrix} = \begin{bmatrix} \lambda_1 v_1^H \\ \vdots \\ \lambda_p v_p^H \end{bmatrix}.$$

Hence, we can write the Spectral Decomposition as

$$A = UDU^H = \begin{bmatrix} v_1 & \dots & v_p \end{bmatrix} \begin{bmatrix} \lambda_1 v_1^H \\ \vdots \\ \lambda_p v_p^H \end{bmatrix}.$$

Entrywise, this is exactly

$$\begin{aligned} A_{ij} &= \sum_{k=1}^{p} U_{ik} \left(DU^H\right)_{kj} \\ &= \sum_{k=1}^{p} (v_k)_i \left(\lambda_k \bar{v}_k\right)_j \\ &= \sum_{k=1}^{p} \lambda_k (v_k)_i \left(\bar{v}_k\right)_j \\ &= \sum_{k=1}^{p} \lambda_k \left(v_k \otimes v_k\right)_{ij} \\ &= \left(\sum_{k=1}^{p} \lambda_k v_k \otimes v_k\right)_{ij}. \end{aligned}$$

Hence, we find

$$\boxed{A = \sum_{k=1}^{p} \lambda_k v_k \otimes v_k = \lambda_1 v_1 \otimes v_1 + \lambda_2 v_2 \otimes v_2 + ... + \lambda_p v_p \otimes v_p} \tag{7.7}$$

Therefore, A can be written as the sum of rank-one (outer-product) matrices, each of which is weighted by the magnitude of an eigenvalue of A.

Alternatively, because the action of the outer product on a vector can be written in terms of the inner product, namely

$$(u \otimes v)x = uv^H x = (v^H x)u = \langle v, x \rangle u,$$

we can use the notation of Theorems 7.17 and 7.18 to find

$$\boxed{Ax = \sum_{k=1}^{p} \lambda_k \langle v_k, x \rangle v_k}$$

for every $x \in \mathbb{C}^p$. Hence, this result is exactly a special case of Theorem 7.18. Regardless, for those more comfortable with matrix algebra we provide an alternate proof below.

Proof of Theorem 7.19. ($\Rightarrow$) Assume A is normal. Then, $A^H A = AA^H$. By Theorem 7.8 A has a Schur decomposition so that there are unitary $U \in \mathbb{C}^{p \times p}$ and upper triangular $T \in \mathbb{C}^{p \times p}$ such that

$$A = UTU^H.$$

Then, taking the Hermitian we find

$$A^H = (UTU^H)^H = UT^H U^H$$

so the normality condition is just

$$UT^H U^H UTU^H = UTU^H UT^H U^H$$

and because $U^H U = \mathbb{I}$, this becomes

$$UT^H TU^H = UTT^H U^H.$$

Since U and U^H are nonsingular, we can left/right multiply by their inverses to arrive at

$$T^H T = TT^H.$$

However, T is upper triangular, which means we can write it as

$$T = \begin{bmatrix} t_{11} & t_{12} & ... & t_{1p} \\ 0 & t_{22} & ... & \vdots \\ \vdots & ... & \ddots & \vdots \\ 0 & ... & 0 & t_{pp} \end{bmatrix}$$

and T^H is given by

$$T^H = \begin{bmatrix} \overline{t_{11}} & 0 & ... & 0 \\ \overline{t_{12}} & \overline{t_{22}} & ... & \vdots \\ \vdots & \dots & \ddots & 0 \\ \overline{t_{1p}} & \dots & \dots & \overline{t_{pp}} \end{bmatrix}.$$

Multiplying these matrices then yields

$$T^H T = \begin{bmatrix} |t_{11}|^2 & \overline{t_{11}}t_{12} & ... & \overline{t_{11}}t_{1p} \\ \overline{t_{12}}t_{11} & |t_{12}|^2 + |t_{22}|^2 & ... & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \overline{t_{1p}}t_{11} & \dots & \dots & \sum_{i=1}^{p} |t_{ip}|^2 \end{bmatrix}$$

and

$$TT^H = \begin{bmatrix} \sum_{j=1}^{p} |t_{1j}|^2 & \dots & \dots & t_{1p}\overline{t_{pp}} \\ \vdots & \sum_{j=2}^{p} |t_{2j}|^2 & \dots & \vdots \\ \vdots & \dots & \ddots & \vdots \\ t_{pp}\overline{t_{1p}} & \dots & \dots & |t_{pp}|^2 \end{bmatrix}.$$

Since $TT^H = T^H T$, we can equate the diagonal entries of these two matrices. Consider the first diagonal entry, which yields the equation

$$\sum_{j=1}^{p} |t_{1j}|^2 = |t_{11}|^2.$$

Upon subtracting the first term in the sum, i.e. $|t_{11}|^2$ from both sides of the equation, we find

$$\sum_{j=2}^{p} |t_{1j}|^2 = 0$$

and since each term is nonnegative, this implies $t_{1j} = 0$ for every $j \geq 2$. Next, we consider the second diagonal entry, which yields the equation

$$\sum_{j=2}^{p} |t_{2j}|^2 = |t_{12}|^2 + |t_{22}|^2.$$

Of course, we have concluded that $|t_{12}|^2 = 0$, which removes this term, and again subtracting the first term in the sum from both sides, we find

$$\sum_{j=3}^{p} |t_{2j}|^2 = 0,$$

which, similar to before, implies $t_{2j} = 0$ for every $j \geq 3$. We successively continue this process for each diagonal entry and find, for instance,

$$|t_{p-1,p-1}|^2 + |t_{p-1,p}|^2 = \sum_{i=1}^{p-1} |t_{i,p-1}|^2$$

which, because $t_{i,p-1} = 0$ for all $i \leq p-2$, implies $|t_{p-1,p}|^2 = 0$, and thus $t_{p-1,p} = 0$. Finally, the last diagonal entry is

$$|t_{pp}|^2 = \sum_{i=1}^{p} |t_{ip}|^2$$

which merely becomes $|t_{pp}|^2 = |t_{pp}|^2$ upon utilizing the previously-determined information that $t_{ip} = 0$ for $i < p$. Thus, from these equations we conclude that $t_{ij} = 0$ for all $i, j = 1, ..., p$ with $i < j$. Hence, T is lower triangular, and since T must also be upper triangular, we conclude that T is actually diagonal. Thus, $A = UDU^H$ where U is unitary and $D = T$ is diagonal, which means A is unitarily diagonalizable.

($\Leftarrow$) Assume $A = UDU^H$ for some unitary $U \in \mathbb{C}^{p\times p}$ and diagonal $D \in \mathbb{C}^{p\times p}$. Then, taking the Hermitian of this representation yields

$$A^H = (UDU^H)^H = UD^HU^H.$$

Computing the product of these matrices and using the fact that $U^HU = \mathbb{I}$, we find

$$A^HA = UD^HU^HUDU^H = UD^HDU^H$$

and

$$AA^H = UDU^HUD^HU^H = UDD^HU^H.$$

However, as diagonal matrices always commute, we see that $D^HD = DD^H$. Therefore, the above representations are equal and we find

$$\begin{aligned} A^HA &= UD^HDU^H \\ &= UDD^HU^H \\ &= AA^H \end{aligned}$$

which is exactly the condition that A is normal. □

Comment. As previously mentioned, Theorems 7.17 and 7.19 are generalizations of Theorem 13 from the Linear Algebra review, which states that $A \in \mathbb{R}^{p\times p}$ is symmetric if and only if A is orthogonally diagonalizable. However, it's important to note that there are real, normal matrices that are NOT symmetric (e.g., orthogonal matrices). Therefore, it is not the case that a normal matrix $B \in \mathbb{R}^{p\times p}$ is automatically orthogonally diagonalizable. Theorem 7.19 guarantees that B is unitarily diagonalizable, but it could certainly possess complex eigenvalues and eigenvectors, which leads to the matrices U and D in the statement of the Spectral Theorem failing to be real. Said another way, B could be a real, normal matrix whose unitary diagonalization is actually complex-valued, and therefore not an orthogonal diagonalization. The following example should suitably demonstrate this idea.

Example 57. Consider $A \in \mathbb{R}^{2\times 2}$ defined by

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

It should be clear that A is real-valued, but not symmetric. Let's check if A is normal - computing

$$A^HA = A^TA = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

and

$$AA^H = AA^T = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

we see that

$$A^H A = AA^H,$$

and A is normal. Because of this, the matrix T in the Schur decomposition of A is diagonal (as seen in the proof of Theorem 7.19), and the diagonal entries of T are merely the eigenvalues of A, as described in Theorem 7.8. Of course, computing the eigenvalues of A, we find

$$\lambda_{1,2} = 1 \pm i$$

and this implies that

$$T = \begin{bmatrix} 1+i & 0 \\ 0 & 1-i \end{bmatrix} \qquad \text{or} \qquad T = \begin{bmatrix} 1-i & 0 \\ 0 & 1+i \end{bmatrix}.$$

Thus, the unitary diagonalization of A guaranteed by the Spectral Theorem is, in fact, complex and not real. This further implies that A is not orthogonally diagonalizable. Of course, we already knew this because A is not symmetric, in view of the theorem from Linear Algebra we previously mentioned (Corollary 7.4 below).

Comment. Such an important theorem deserves a few additional comments:

1. Just as in the previous diagonalizations we have encountered, the diagonal entries of D are exactly the eigenvalues of A.
2. Additionally, U can be chosen so that the eigenvalues of A appear on the diagonal of D in any desired order.
3. As a side note, notice that within the ($\Rightarrow$) direction of the proof, we've actually shown that if $T \in \mathbb{C}^{p\times p}$ is upper triangular and normal, then T must be diagonal. The same is true if T is lower triangular and normal.

Corollary 7.4. Let $A \in \mathbb{R}^{p\times p}$ be given. Then, the following statements are equivalent

1. A is normal and possesses only real eigenvalues
2. A is symmetric
3. A is orthogonally diagonalizable

We do not include a proof, but note that it is similar to the proof of Theorem 7.19 with Hermitian operations replaced by transposes. Curious readers will find a proof in [11], though we include an example for completeness.

Example 58. Consider $A \in \mathbb{R}^{3\times 3}$ defined by

$$A = \begin{bmatrix} 0 & 2 & -1 \\ 2 & 3 & -2 \\ -1 & -2 & 0 \end{bmatrix}.$$

Notice that A is symmetric, so by Corollary 7.4, it should be orthogonally diagonalizable. To show this, we use the following algorithm:

1. Compute the eigenvalues and associated eigenspaces of A.

   From the standard techniques of Linear Algebra (i.e., finding $\lambda \in \mathbb{R}$ such that $\det(A - \lambda \mathbb{I}) = 0$ and then computing $v \in \mathbb{R}^3 \setminus \{0\}$ such that $Av = \lambda v$ for each λ), we find only two eigenvalues

   $$\lambda_1 = -1 \qquad \text{and} \qquad \lambda_2 = 5$$

   with the associated eigenspaces

   $$\mathcal{E}_1 = \left\{ \alpha \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix} + \beta \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} : \ \alpha, \beta \in \mathbb{R} \right\} \quad \text{and} \quad \mathcal{E}_2 = \left\{ \gamma \begin{bmatrix} -1 \\ -2 \\ 1 \end{bmatrix} : \ \gamma \in \mathbb{R} \right\}.$$

2. Use Gram-Schmidt (if necessary) to compute an orthonormal basis for $\mathbb{R}^3$ from the eigenspaces.

   Utilizing Gram-Schmidt in our example yields the orthonormal basis

   $$B = \left\{ \begin{bmatrix} \frac{1}{\sqrt{2}} \\ 0 \\ \frac{1}{\sqrt{2}} \end{bmatrix}, \begin{bmatrix} -\frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \\ \frac{1}{\sqrt{3}} \end{bmatrix}, \begin{bmatrix} -\frac{1}{\sqrt{6}} \\ -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{bmatrix} \right\}.$$

3. From the eigenvalues, form the diagonal matrix $D = \operatorname{diag}(\lambda_1, \lambda_2, \lambda_3)$ with repetition.

   In our example, this is merely

   $$D = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 5 \end{bmatrix}.$$

4. Form the orthogonal matrix P with columns $\{v_1, ..., v_p\}$ from the orthonormal basis, where the columns of P are grouped according to the eigenspace from which they arose.

   Since the eigenvalue $\lambda_1 = -1$ appears in the first two diagonal entries of D, we use the orthonormal vectors arising from $\mathcal{E}_1$ within the first two columns of P. Therefore,

   $$P = \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{1}{\sqrt{3}} & -\frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{6}} \end{bmatrix}$$

   and we finally have $A = PDP^T$.

The outer-product representation (7.7) of a normal or symmetric matrix is especially important if A has a **spectral gap** (i..e., a large difference in magnitude between consecutive eigenvalues when ordered), as it may allow us to truncate this sum and store a strong approximation of A using a significantly smaller amount of data. This is a topic that will be explored in greater detail within the next chapter.

7.8 Singular Value Decomposition

Finally, we come to one of the more important theorems and constructions in all of the mathematical sciences, the Singular Value Decomposition. This fundamental matrix decomposition is widely used throughout engineering and the sciences to solve numerous applied problems. Applications of the SVD include image compression (and more generally data compression), least-squares solutions of linear systems, low-rank matrix approximations, curve fitting and linear regression, pattern recognition, filtering, signal processing, shape optimization, and sensitivity analysis, amongst others. Such processes are utilized within a variety fields, for instance to find patterns in statistical data, predict weather using atmospheric models, compare the structure of molecules, compute the spatiotemporal trajectories of epidemics, and even search for gravitational waves. Though some of the aforementioned operations can be performed using other means, the SVD is often the tool of choice for these applications because numerical methods have been developed to accurately and efficiently approximate the SVD to a high degree of precision. It's mathematical importance also cannot be overstated as it provides a decomposition of any matrix in terms of associated orthonormal vectors, and this topic is further explored in the celebrated expository article [7].

Theorem 7.20 (Singular Value Decomposition)**.** Let $\mathcal{V}$ be a Hilbert space with an orthonormal basis and $T : \mathcal{V} \to \mathcal{V}$ be a compact linear operator. Then, there exist orthonormal sequences $\{u_n\}_{n=1}^{\infty}$ and $\{v_n\}_{n=1}^{\infty}$ of $\mathcal{V}$ such that

$$T(v) = \sum_{n=1}^{\infty} \sigma_n \langle v_n, v \rangle u_n$$

for every $v \in \mathcal{V}$, where $\sigma_n \in \mathbb{R}$ is a sequence of positive numbers. Furthermore, if σ_n has infinitely many terms, then $\sigma_n \to 0$ as $n \to \infty$.

Though we do not include a proof here, this can be found in [9]. We will instead prove the finite-dimensional version of this result, which contains nearly all of the same ideas and techniques.

Comment. Notice that T is not assumed to be normal in Theorem 7.20. So, in the finite-dimensional case, this decomposition can be performed on any matrix $A \in \mathbb{C}^{p\times q}$ or $A \in \mathbb{R}^{p\times q}$.

Theorem 7.21 (Singular Value Decomposition for $\mathbb{C}^{p\times q}$)**.** Let $A \in \mathbb{C}^{p\times q}$ be given. Then, there are unitary matrices $U \in \mathbb{C}^{p\times p}$ and $V \in \mathbb{C}^{q\times q}$ and a matrix $\Sigma \in \mathbb{R}^{p\times q}$ of the form

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & 0 & \dots \\ \dots & 0 & \ddots & \dots \\ 0 & \dots & 0 & \sigma_q \\ 0 & \dots & \dots & 0 \end{bmatrix}$$

with

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > \sigma_{k+1} = \dots = \sigma_q = 0$$

for some $0 \leq k \leq q$ such that

$$\boxed{A = U\Sigma V^H.}$$

In this case, the real numbers $\sigma_1, ..., \sigma_k > 0$ are called the **singular values** of A.

If, instead, $A \in \mathbb{R}^{p\times q}$, then $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{q\times q}$ are orthogonal rather than unitary, and the exact same decomposition holds with $V^H = V^T$ so that

$$\boxed{A = U\Sigma V^T.}$$

Comment. Though we write only $k \leq q$ in the theorem, from the matrix representation of Σ above, one can see that, in fact, $k \leq \min\{p, q\}$.

Proof. Just as the finite-dimensional Spectral Theorem (Theorem 7.19) was just a corollary of Theorem 7.18, this result is merely a corollary of Theorem 7.20 under the exact same circumstances. We apply this theorem to the Hilbert space $\mathcal{V} = \mathbb{C}^p$ with inner product

$$\langle x, y\rangle = x^H y = \sum_{n=1}^{p} \bar{x}_k y_k$$

and orthonormal basis

$$B = \{e_k : k = 1, ..., p\},$$

and $T : \mathbb{C}^p \to \mathbb{C}^p$ defined by

$$T(x) = Ax$$

for all $x \in \mathbb{C}^p$ (where A is the given matrix) is a compact normal operator, this result follows as an immediate corollary of Theorem 7.20.

As before, for those more comfortable with matrix algebra, we provide a direct proof of the result for $A \in \mathbb{R}^{p\times q}$, which uses nearly every tool we've discussed in previous chapters. Just a brief warning - this will be a lengthy proof.

We will prove the result for $p \geq q$. If instead, $p \leq q$, then we merely apply the forthcoming proof to A^T instead. If $A = 0$, then $\Sigma = 0$ with U and V arbitrary orthogonal matrices, and the result follows with $k = 0$. Otherwise, we can conclude that $A^T A$ has at least one nonzero eigenvalue.

Now, since $A \in \mathbb{R}^{p\times q}$, we find $A^T A \in \mathbb{R}^{q\times q}$ and because $\left(A^T A\right)^T = A^T A$ this matrix is symmetric. By the Spectral Theorem for symmetric matrices (i.e., Corollary 7.4), there exists an orthogonal $V \in \mathbb{R}^{q\times q}$ and a diagonal $D \in \mathbb{R}^{q\times q}$ such that

$$A^T A = VDV^T.$$

Let $\lambda_1, ..., \lambda_q$ be the diagonal entries of D and $v_1, ..., v_q$ be the columns of V. Recall from the spectral decomposition of Corollary 7.4 that the λ_j merely represent eigenvalues of $A^T A$ while the $v_j \in \mathbb{R}^q$ are their corresponding eigenvectors so that

$$A^T A v_j = \lambda_j v_j \tag{7.8}$$

for all $j = 1, ..., q$. Additionally, because V is an orthogonal matrix, the set $\{v_1, ..., v_q\}$ is orthonormal.

Next, we notice that $\lambda_1, ..., \lambda_q \geq 0$. Indeed, beginning with the equation $A^TAv = \lambda v$ and multiplying on the left by v^T, we find

$$v^TA^TAv = v^T\lambda v$$

and therefore

$$(Av)^TAv = \lambda v^Tv.$$

Finally, this is equivalent to

$$\|Av\|_2^2 = \lambda\|v\|_2^2,$$

and upon dividing by $\|v\|_2^2 \neq 0$, we have an expression for λ, namely

$$\lambda = \frac{\|Av\|_2^2}{\|v\|_2^2} \geq 0.$$

Since (λ, v) is an arbitrary eigenpair, we see that all eigenvalues are nonnegative, i.e., $\lambda_1, ..., \lambda_q \geq 0$. Thus, we order these scalars and define k to be the number of nonzero eigenvalues of A^TA so that

$$\lambda_1 \geq \lambda_2 \geq ... \geq \lambda_k > \lambda_{k+1} = ... = \lambda_q = 0.$$

With this, we define $\sigma_j = \sqrt{\lambda_j}$ for $j = 1, ..., q$ and note

$$\sigma_1 \geq \sigma_2 \geq ... \geq \sigma_k > \sigma_{k+1} = ... = \sigma_q = 0.$$

Further define the vectors $u_j \in \mathbb{R}^p$ by

$$u_j = \frac{1}{\sigma_j}Av_j$$

for all $j = 1, ..., k$, and note that $k \leq q$.

We will show that the set $\{u_1, ..., u_k\}$ is, in fact, orthonormal. Indeed, recalling the Kronecker delta function

$$\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & i \neq j \end{cases}$$

and using (7.8) to compute the standard inner product yields

$$\begin{aligned} u_i^Tu_j &= \frac{1}{\sigma_i}(Av_i)^T\frac{1}{\sigma_j}(Av_j) \\ &= \frac{1}{\sigma_i\sigma_j}v_i^TA^TAv_j \\ &= \frac{1}{\sigma_i\sigma_j}v_i^T\lambda_jv_j \\ &= \frac{\lambda_j}{\sigma_i\sigma_j}v_i^Tv_j \\ &= \frac{\lambda_j}{\sqrt{\lambda_i}\sqrt{\lambda_j}}\delta_{ij} \\ &= \delta_{ij} \end{aligned}$$

for any $i, j = 1, ..., k$. Hence, $u_i^T u_j = 0$ if $i \neq j$, while $u_i^T u_j = 1$ if $i = j$. Thus, $\{u_1, ..., u_k\}$ is orthonormal.

Of course, we've only defined σ_j and u_j for $j = 1, ..., k$. Since $p \geq q$, it follows that $k \leq q \leq p$. So, we may need to construct additional vectors (i.e., u_j for $j = k+1, ..., p$) to represent columns of $U \in \mathbb{R}^{p \times p}$. Therefore, we will find $p - k$ additional orthonormal vectors, i.e. $u_{k+1}, ..., u_p$ that satisfy the condition $u_i^T u_j = 0$ for all $i, j = 1, ..., p$ with $i \neq j$. To this end, define $B \in \mathbb{R}^{k \times p}$ by

$$B = \begin{bmatrix} u_1^T \\ \vdots \\ u_k^T \end{bmatrix}.$$

Notice that, by construction, any vector in $\text{Nul}(B)$ is necessarily orthogonal to each of the vectors $u_1, ..., u_k$. By the Rank-Nullity Theorem (Theorem 5.3), we find

$$\text{rank}(B) + \dim(\text{Nul}(B)) = p.$$

Additionally, $\text{rank}(B) = \text{rank}(B^T) = k$ because the columns of B^T consist of k orthogonal (and hence linearly independent) vectors. Thus, we conclude

$$\dim(\text{Nul}(B)) = p - \text{rank}(B) = p - k.$$

So, there exists a basis of $p - k$ vectors for $\text{Nul}(B)$. Using the Gram-Schmidt process and normalization, we construct an orthonormal basis for $\text{Nul}(B)$ from these $p - k$ vectors and define $\{u_{k+1}, ..., u_p\}$ to be these newly-created vectors in the orthonormal basis.

Finally, define $U \in \mathbb{R}^{p \times p}$ by

$$U = \begin{bmatrix} u_1 & \dots & u_p \end{bmatrix}$$

and $\Sigma \in \mathbb{R}^{p \times q}$ by

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & ... & 0 \\ 0 & \sigma_2 & 0 & ... \\ ... & 0 & \ddots & ... \\ 0 & ... & 0 & \sigma_q \\ 0 & ... & ... & 0 \end{bmatrix}.$$

Because $V \in \mathbb{R}^{q \times q}$ has already been defined, we merely need to verify the relationship $A = U\Sigma V^T$. Previously, we saw that $Av_j = \sigma_j u_j$ for all $j = 1, ..., k$ by definition. For $j = k+1, ..., q$, we know that $\sigma_j = 0$ and $\lambda_j = 0$; hence, we find

$$\begin{aligned} \|Av_j\|_2^2 &= (Av_j)^T (Av_j) \\ &= v_j^T A^T A v_j \\ &= v_j^T \lambda_j v_j \\ &= \lambda_j \|v_j\|_2^2. \end{aligned}$$

Because $\lambda_j = 0$, this implies $Av_j = 0$, and so $Av_j = \sigma_j u_j$ for $j = k+1, ..., q$ as well. Therefore, $Av_j = \sigma_j u_j$ for $j = 1, ..., q$, and this is exactly

$$(AV)_{j\text{th column}} = (U\Sigma)_{j\text{th column}}$$

for all $j = 1, ..., q$, and thus

$$AV = U\Sigma.$$

Since V is orthogonal (and thus $V^T = V^{-1}$) multiplying by V^T on the right yields

$$A = U\Sigma V^T$$

and the proof is complete. □

Comment. Though our proof assumes $p \geq q$ and uses A^TA to construct the SVD, we could just as easily have assumed $p \leq q$ and replaced A with A^T everywhere. Of course, doing this means we would have used the matrix

$$(A^T)^T(A^T) = AA^T$$

to construct the SVD. This is not merely a symptom of the proof. When computing the SVD by hand for small examples, it is easiest to use A^TA when $p \geq q$ (i.e., when A is a "tall" matrix) and AA^T when $p \leq q$ (i.e., when A is a "wide" matrix). Notice that in each case $A^TA \in \mathbb{R}^{q\times q}$ and $AA^T \in \mathbb{R}^{p\times p}$. So, when $p \geq q$, we'd rather compute the smaller matrix $A^TA \in \mathbb{R}^{q\times q}$, while when $p \leq q$, we'd again prefer to compute the smaller matrix, but this time it's $AA^T \in \mathbb{R}^{p\times p}$.

Comment. Similar to the orthogonal diagonalization of symmetric matrices, the SVD can be expressed as a sum of outer products, namely

$$A = \sum_{j=1}^{k} \sigma_j u_j \otimes v_j$$

or equivalently

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + ... + \sigma_k u_k v_k^T$$

where k is the number of nonzero singular values. Of course, we could also define $k = \text{rank}(A)$, as well. This form of the SVD will be useful for constructing truncated approximations of A and implementing Principal Component Analysis in the next chapter.

Example 59. Define $A \in \mathbb{R}^{3\times 2}$ by

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

We wish to compute the SVD of A. Since $p \geq q$, we begin by computing

$$A^TA = \begin{bmatrix} 1 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 2 & 2 \end{bmatrix} = \begin{bmatrix} 9 & 9 \\ 9 & 9 \end{bmatrix}.$$

The eigenvalues of this matrix are $\lambda_1 = 18$ and $\lambda_2 = 0$ with corresponding eigenvectors

$$x_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad x_2 = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

These vectors are orthogonal, and we normalize them so that

$$v_1 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \qquad \text{and} \qquad v_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} -1 \\ 1 \end{bmatrix}$$

form the columns of the orthogonal matrix

$$V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

From the eigenvalues, we see that $k = 1$ (i.e., there is just one nonzero eigenvalue of $A^T A$) and define $\sigma_1 = \sqrt{18}$ and $\sigma_2 = 0$, then form the corresponding matrix

$$\Sigma = \begin{bmatrix} \sqrt{18} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

Next, we define the first column of the U matrix by

$$u_1 = \frac{1}{\sqrt{18}} A v_1 = \frac{1}{\sqrt{18}} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \frac{1}{6} \begin{bmatrix} 2 \\ 4 \\ 4 \end{bmatrix} = \frac{1}{3} \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}.$$

Since $U \in \mathbb{R}^{3\times 3}$, we still need to construct u_2 and u_3. Let

$$B = u_1^T = \frac{1}{3} \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}$$

so that

$$\text{Nul}(B) = \left\{ x \in \mathbb{R}^3 : \ \frac{1}{3}x_1 + \frac{2}{3}x_2 + \frac{2}{3}x_3 = 0 \right\}.$$

Hence, for any $x \in \text{Nul}(B)$, the components of x satisfy

$$x_1 = -2x_2 - 2x_3.$$

Of course, we can use a parametric representation so that $x \in \text{Nul}(B)$ implies

$$x = \begin{bmatrix} -2x_2 - 2x_3 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} x_2 + \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} x_3.$$

Thus, the two vectors

$$y_1 = \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad y_2 = \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix}$$

form a basis for $\text{Nul}(B)$ and are both orthogonal to u_1. Unfortunately, they're not orthogonal to each other, and we must use Gram-Schmidt to construct an orthogonal basis for this subspace. So, let $a = y_1$ and compute

$$\begin{aligned} b &= y_2 - \frac{y_1^T y_2}{\|y_1\|_2^2} y_1 \\ &= \begin{bmatrix} -2 \\ 0 \\ 1 \end{bmatrix} - \frac{4}{5} \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} \\ &= \frac{1}{5} \begin{bmatrix} -2 \\ -4 \\ 5 \end{bmatrix} \end{aligned}$$

Finally, we define

$$u_2 = \frac{a}{\|a\|_2} = \frac{1}{\sqrt{5}} \begin{bmatrix} -2 \\ 1 \\ 0 \end{bmatrix} \qquad \text{and} \qquad u_3 = \frac{b}{\|b\|_2} = \frac{1}{3\sqrt{5}} \begin{bmatrix} -2 \\ -4 \\ 5 \end{bmatrix}$$

and therefore

$$U = \begin{bmatrix} \frac{1}{3} & -\frac{2}{\sqrt{5}} & -\frac{2}{3\sqrt{5}} \\ \frac{2}{3} & \frac{1}{\sqrt{5}} & -\frac{4}{3\sqrt{5}} \\ \frac{2}{3} & 0 & \frac{5}{3\sqrt{5}} \end{bmatrix}.$$

With this, we have constructed the decomposition $A = U\Sigma V^T$ where these matrices are defined above.

7.9 Properties and Applications of SVD

With the fundamentals of the SVD established in the previous section, we now turn our attention to proving some of its more interesting and useful properties for matrices. These include theoretical properties, computational properties, and specific characteristics that can be used in a variety of applications. In particular, we will study the following:

1. Orthonormal bases for the four fundamental subspaces of A
2. Computing rank using the SVD
3. Computing matrix norms using the SVD
4. Least squares solutions using the SVD
5. The Moore-Penrose Pseudoinverse

First, we note that - by its very construction - the Singular Value Decomposition of a matrix $A \in \mathbb{R}^{p \times q}$ inherently provides orthonormal bases for the four fundamental subspaces generated by the matrix, namely $\text{Nul}(A), \text{Col}(A^T) \subseteq \mathbb{R}^p$ and $\text{Nul}(A^T), \text{Col}(A) \subseteq \mathbb{R}^q$.

Theorem 7.22. Let $A \in \mathbb{R}^{p\times q}$ be given with SVD $A = U\Sigma V^T$ and k denoting the number of nonzero singular values of A. Then, we have

1. $\mathcal{U}_1 := \{u_1, ..., u_k\}$ is an orthonormal basis for $\text{Col}(A)$
2. $\mathcal{U}_2 := \{u_{k+1}, ..., u_p\}$ is an orthonormal basis for $\text{Nul}(A^T)$
3. $\mathcal{V}_1 := \{v_1, ..., v_k\}$ is an orthonormal basis for $\text{Col}(A^T)$
4. $\mathcal{V}_2 := \{v_{k+1}, ..., v_q\}$ is an orthonormal basis for $\text{Nul}(A)$.

Proof. Recalling the proof of Theorem 7.21, we merely notice that $u_j \in \mathbb{R}^p$ are orthonormal vectors satisfying

$$u_j = \frac{1}{\sigma_j} Av_j = Aw_j$$

where $w_j = \frac{1}{\sigma_j} v_j$ for all $j = 1, ..., k$. Hence, $u_j \in \text{Col}(A)$ for every $j = 1, ..., k$. By definition, we see that $\mathcal{U}_1$ represents an orthonormal set of k vectors in $\text{Col}(A)$, and because $k = \text{rank}(A) = \dim(\text{Col}(A))$, it follows that these k vectors must form a basis for $\text{Col}(A)$. Similarly, $u_{k+1}, ..., u_p$ are defined to be an orthonormal basis for $\text{Nul}(B)$ where B is the matrix whose rows consist of $u_1, ..., u_k$. Thus, each of $u_{k+1}, ..., u_p$ are orthogonal to these k vectors, which means they are in $\text{Col}(A)^\perp$. Of course, we know from the Fundamental Theorem of Linear Algebra that $\text{Col}(A)^\perp = \text{Nul}(A^T)$, and using the Rank-Nullity Theorem, we see that $\dim(\text{Nul}(A^T)) = p - k$. As $\mathcal{U}_2$ is a set of $p - k$ orthonormal vectors in $\text{Nul}(A^T)$, it must be a basis for this subspace.

Turning to $\mathcal{V}_1$ and $\mathcal{V}_2$, we note that by (7.8) we have

$$A^T Av_j = \lambda_j v_j$$

for every $j = 1, ..., q$. In particular, for $j = 1, ..., k$ we know that $\lambda_j = \sigma_j^2 \neq 0$ so we can express v_j as

$$v_j = \frac{1}{\lambda_j} A^T Av_j = A^T w_j$$

where $w_j = \frac{1}{\lambda_j} Av_j$. This shows that $v_j \in \text{Col}(A^T)$ for every $j = 1, .., k$. Additionally, we know

$$\dim(\text{Col}(A^T)) = \text{rank}(A^T) = \text{rank}(A) = k,$$

and because $\mathcal{V}_1$ consists of k orthonormal vectors in $\text{Col}(A^T)$, this set must form a basis for the subspace. Finally, for $j = k + 1, ..., q$ we see that $\lambda_j = \sigma_j^2 = 0$ so v_j satisfies

$$A^T Av_j = 0.$$

Hence, $v_j \in \text{Nul}(A^T A)$ for every $j = k + 1, ..., q$. As demonstrated by a previous homework problem, $\text{Nul}(A^T A) = \text{Nul}(A)$, and therefore $v_j \in \text{Nul}(A)$ for every $j = k + 1, ..., q$. Again invoking the Rank-Nullity theorem, we find

$$\dim(\text{Nul}(A)) = q - \text{rank}(A) = q - k.$$

As $\mathcal{V}_2$ consists of $q - k$ orthonormal vectors in $\text{Nul}(A)$, they must form a basis for this subspace, and this completes the proof. □

Next, we consider the task of computing the rank of a given matrix.

Theorem 7.23. Let $A \in \mathbb{R}^{p\times q}$ be given with SVD $A = U\Sigma V^T$ and k denoting the number of nonzero singular values of A. Then,

$$\text{rank}(A) = \text{rank}(\Sigma) = k.$$

Proof. Recall that by Theorem 2.6, we have the result

$$\text{rank}(BC) \leq \min\{\text{rank}(B), \text{rank}(C)\}$$

for any $B \in \mathbb{R}^{p\times q}$ and $C \in \mathbb{R}^{q\times r}$. Therefore, we find

$$\text{rank}(A) = \text{rank}(U\Sigma V^T) \leq \text{rank}(\Sigma V^T) \leq \text{rank}(\Sigma) = k$$

where the last equality follows from the form of Σ. Of course, since U and V are orthogonal we can invert the SVD relationship to solve for Σ to find

$$\Sigma = U^T AV$$

and thus

$$\text{rank}(\Sigma) = \text{rank}(U^T AV) \leq \text{rank}(AV) \leq \text{rank}(A).$$

Combining these two inequalities yields

$$\text{rank}(A) = \text{rank}(\Sigma) = k$$

and the result follows. □

Comment. In computing the SVD, we automatically determine the rank of the given matrix A. Hence, assuming that algorithms exist to quickly and efficiently compute the SVD (and they do), this number can be well-approximated with ease. In fact, the Matlab command `rank(A)` computes the rank of A as the number of singular values of A that are larger than a specified (or default), small tolerance.

Next, we consider the problem of computing norms of a matrix.

Theorem 7.24. Let $A \in \mathbb{R}^{p\times q}$ be given with SVD $A = U\Sigma V^T$ and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > \sigma_{k+1} = \ldots = \sigma_q = 0$$

denoting the ordered singular values of A. Then,

$$\|A\|_2 = \|\Sigma\|_2 = \sigma_1.$$

Proof. Recall the definition of this norm, namely

$$\|A\|_2 = \max_{x\in\mathbb{R}^q\setminus\{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x\in\mathbb{R}^q\setminus\{0\}} \frac{\|U\Sigma V^T x\|_2}{\|x\|_2}.$$

Now, by Theorem 7.5 (more specifically, the comment after the theorem), we know that left multiplication of a vector by an orthogonal matrix will preserve the norm of the original vector. Therefore, we have

$$\|U\Sigma V^T x\|_2 = \|\Sigma V^T x\|_2$$

and the computation of $\|A\|_2$ is simplified.

Next, for any $x \in \mathbb{R}^q \setminus \{0\}$, we let $y = V^T x$ and note that because V^T is nonsingular, $x = 0$ if and only if $y = 0$. Theorem 7.7 guarantees that V^T is orthogonal because V orthogonal, and thus

$$\|x\|_2 = \|Vy\|_2 = \|y\|_2$$

where we have again used Theorem 7.5 to establish the last equality. Therefore, we find

$$\|A\|_2 = \max_{x \in \mathbb{R}^q \setminus \{0\}} \frac{\|\Sigma V^T x\|_2}{\|x\|_2} = \max_{y \in \mathbb{R}^q \setminus \{0\}} \frac{\|\Sigma y\|_2}{\|y\|_2} = \|\Sigma\|_2.$$

Thus, the first portion of the theorem has been shown, and we focus on proving $\|\Sigma\|_2 = \sigma_1$ by showing that this quantity is both greater than and less than σ_1.

We will first show that $\|\Sigma\|_2 \leq \sigma_1$. Notice that for any $y \in \mathbb{R}^q$, we have

$$\|\Sigma y\|_2^2 = (\Sigma y)^T (\Sigma y) = \begin{bmatrix} \sigma_1 y_1 & \dots & \sigma_k y_k & 0 & \dots & 0 \end{bmatrix} \begin{bmatrix} \sigma_1 y_1 \\ \vdots \\ \sigma_k y_k \\ 0 \\ \vdots \\ 0 \end{bmatrix} = \sum_{j=1}^{k} \sigma_j^2 y_j^2.$$

Hence, because the singular values are ordered this further implies

$$\sum_{j=1}^{k} \sigma_j^2 y_j^2 \leq \sigma_1^2 \sum_{j=1}^{k} y_j^2 \leq \sigma_1^2 \|y\|_2^2$$

so that combining these inequalities produces

$$\|\Sigma y\|_2^2 \leq \sigma_1^2 \|y\|_2^2.$$

Upon dividing by $\|y\|_2^2$ and taking the square root, this becomes

$$\frac{\|\Sigma y\|_2}{\|y\|_2} \leq \sigma_1$$

for any $y \in \mathbb{R}^q$. Finally, taking the maximum over all such vectors $y \neq 0$ yields

$$\|\Sigma\|_2 \leq \sigma_1. \tag{7.9}$$

For the opposite inequality, we need only find one vector that works. So, let $z = e_1 \in \mathbb{R}^q$. Then, $\|z\|_2 = 1$ and thus we find

$$\|\Sigma z\|_2^2 = \sigma_1^2 z_1^2 = \sigma_1^2 = \sigma_1^2 \|z\|_2^2.$$

As before, this becomes

$$\frac{\|\Sigma z\|_2}{\|z\|_2} = \sigma_1$$

and therefore

$$\|\Sigma\|_2 = \max_{x \in \mathbb{R}^q \setminus \{0\}} \frac{\|\Sigma x\|_2}{\|x\|_2} \geq \frac{\|\Sigma z\|_2}{\|z\|_2} = \sigma_1.$$

Pairing this inequality with (7.9) yields $\|\Sigma\|_2 = \sigma_1$ and completes the proof. □

In addition to the SVD allowing us to compute the 2-norm, a similar result holds for the Frobenius norm. First, however, we'll need a lemma prior to establishing this result, namely that the Frobenius norm is invariant (i.e. unchanged) by rotation matrices.

Lemma 7.25. Let $A \in \mathbb{R}^{p\times q}$ be given and assume $U \in \mathbb{R}^{p\times p}$ and $V \in \mathbb{R}^{q\times q}$ are orthogonal matrices. Then, we have

$$\|UA\|_F = \|A\|_F \qquad \text{and} \qquad \|AV\|_F = \|A\|_F.$$

Proof. First, recall the vector definition of $\|\cdot\|_F$, namely

$$\|B\|_F = \sqrt{\sum_{j=1}^{q} \|b_j\|_2^2}$$

for any $B \in \mathbb{R}^{p\times q}$, where $b_1, ..., b_q$ are the columns of B. Denote the columns of A by $a_1, ..., a_q$. Using the second conclusion of Theorem 7.5 applied to $T(v) = Uv$, we can write

$$\begin{aligned}
\|UA\|_F &= \left\| \begin{bmatrix} Ua_1 & \dots & Ua_q \end{bmatrix} \right\|_F \\
&= \sqrt{\sum_{j=1}^{q} \|Ua_j\|_2^2} \\
&= \sqrt{\sum_{j=1}^{q} \|a_j\|_2^2} \\
&= \|A\|_F.
\end{aligned}$$

To prove the second conclusion, we first note that $\|B^T\|_F = \|B\|_F$ for any $B \in \mathbb{R}^{p\times q}$, which follows from the entrywise definition of the Frobenius norm, namely

$$\|B\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q} |b_{ij}|^2} = \sqrt{\sum_{j=1}^{q}\sum_{i=1}^{p} |b_{ji}|^2} = \|B^T\|_F.$$

With this, the result follows by expressing AV as $(V^TA^T)^T$, noting that V^T is orthogonal, and using the first conclusion of the theorem so that

$$\|AV\|_F = \|(V^TA^T)^T\|_F = \|V^TA^T\|_F = \|A^T\|_F = \|A\|_F.$$

□

Theorem 7.26. Let $A \in \mathbb{R}^{p\times q}$ be given with SVD $A = U\Sigma V^T$ and

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k > \sigma_{k+1} = ... = \sigma_q = 0$$

denoting the ordered singular values of A. Then,

$$\|A\|_F = \|\Sigma\|_F = \sqrt{\sum_{j=1}^{k} \sigma_j^2}.$$

Proof. From Lemma 7.25, we can easily write

$$\|A\|_F = \|U\Sigma V^T\|_F = \|\Sigma V^T\|_F = \|\Sigma\|_F$$

because U and V^T are orthogonal. Furthermore, writing the definition of the Frobenius norm yields

$$\|\Sigma\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{q}|\sigma_{ij}|^2} = \sqrt{\sum_{\ell=1}^{k}\sigma_\ell^2}$$

since $\sigma_1, ..., \sigma_k$ are the only nonzero entries of Σ, where σ_{ij} represents the (i,j)th entry of Σ and σ_ℓ represents the ℓth singular value of Σ. □

Comment. Similar to the rank of a matrix, the SVD allows us to compute the 2-norm and Frobenius norm of a given matrix very quickly. In fact, we merely need to compute the first (greatest) singular value in order to calculate the former. The Matlab commands `norm(A)` and `norm(A, 'fro')` compute $\|A\|_2$ as the greatest singular value of A and $\|A\|_F$ as the root of the sum of the squares of all singular values of A, respectively.

Another nice property of the SVD is that it allows us to compute least squares solutions very easily. In fact, we can write down an explicit formula for such a solution merely in terms of the components of the SVD.

Theorem 7.27. Given $A \in \mathbb{R}^{p\times q}$ with singular value decomposition $A = U\Sigma V^T$ and $b \in \mathbb{R}^p$, the least squares solution of $Ax = b$ possessing minimal $\|\cdot\|_2$ norm is exactly the vector

$$x_{\min} = \sum_{j=1}^{k}\frac{c_j}{\sigma_j}v_j$$

where $c = U^Tb$, v_j is the jth column of V, σ_j is the jth singular value of A, and k is the number of nonzero singular values of A.

Proof. To prove the result, we'll establish that

$$\|Ax_{\min} - b\|_2 \leq \|Ax - b\|_2$$

for every $x \in \mathbb{R}^q$. So, given $A = U\Sigma V^T$, $b \in \mathbb{R}^p$, and any $x \in \mathbb{R}^q$, we first define the vectors

$$y = V^Tx \in \mathbb{R}^q \qquad \text{and} \qquad c = U^Tb \in \mathbb{R}^p.$$

Then, using the SVD, the fact that U is orthogonal, and Theorem 7.5, we compute

$$\begin{aligned}\|Ax - b\|_2 &= \|U\Sigma V^Tx - UU^Tb\|_2\\ &= \|U\left(\Sigma V^Tx - U^Tb\right)\|_2\\ &= \|\Sigma V^Tx - U^Tb\|_2\\ &= \|\Sigma y - c\|_2.\end{aligned}$$

Since $y = V^Tx$ and V is nonsingular, we see that $y = 0$ if and only if $x = 0$. Furthermore, because both y and c are merely rotations of the vectors x and b, we

conclude that $x \in \mathbb{R}^q$ minimizes $\|Ax - b\|_2$ if and only if $y \in \mathbb{R}^q$ minimizes $\|\Sigma y - c\|_2$.

Taking a deeper look at the latter norm, we can rewrite this quantity and use the structure of Σ to find

$$(\Sigma y)_i = \sigma_i y_i$$

for every $i = 1, ..., k$ so that

$$\|\Sigma y - c\|_2 = \sqrt{\sum_{i=1}^{p} (\sigma_i y_i - c_i)^2} = \sqrt{\sum_{i=1}^{k} (\sigma_i y_i - c_i)^2 + \sum_{i=k+1}^{p} c_i^2}$$

since $\sigma_{k+1} = \cdots = \sigma_p = 0$. Since $c = U^T b$ is fixed, we have no control over the c_i terms. Thus, minimizing the quantity on the right side is equivalent to the condition

$$y_i = \frac{c_i}{\sigma_i} \tag{7.10}$$

for every $i = 1, ..., k$. Of course, (7.10) is independent of $y_{k+1}, .., y_q \in \mathbb{R}$ so there may be infinitely many such vectors $y \in \mathbb{R}^q$ satisfying this condition if $k \neq q$. Inverting the relationship between x and y, we find $x = Vy$ so that

$$x = \sum_{j=1}^{q} y_j v_j$$

where v_j is the jth column of V, and thus (7.10) is equivalent to

$$x = \sum_{j=1}^{k} \frac{c_j}{\sigma_j} v_j + \sum_{j=k+1}^{q} y_j v_j. \tag{7.11}$$

Therefore, any least squares solution is of this form. Of course, choosing $y_{k+1} = \cdots = y_q = 0$ yields exactly the definition of $x_{\min}$. Hence, $x_{\min}$ is a least squares solution of $Ax = b$.

It remains to show that $x_{\min}$ actually minimizes $\|\cdot\|_2$ among all such least squares solutions. Indeed, computing the norm for any least squares solution we find

$$\|x\|_2 = \|Vy\|_2 = \|y\|_2 = \sqrt{\sum_{j=1}^{q} y_j^2} = \sqrt{\sum_{j=1}^{k} \left|\frac{c_j}{\sigma_j}\right|^2 + \sum_{j=k+1}^{q} y_j^2}.$$

Since the last terms are all nonnegative and $y_{k+1}, .., y_q \in \mathbb{R}$ are arbitrary in choosing a least squares solution, we can minimize this quantity by choosing them all to be zero. Of course, this choice of y yields exactly $x = x_{\min}$, which demonstrates that $x_{\min}$ is the minimal (in $\|\cdot\|_2$) least squares solution. □

Example 60. Define $A \in \mathbb{R}^{3\times 2}$ and $b \in \mathbb{R}^3$ by

$$A = \begin{bmatrix} 1 & 1 \\ 2 & 2 \\ 2 & 2 \end{bmatrix} \qquad \text{and} \qquad b = \begin{bmatrix} 15 \\ 15 \\ -30 \end{bmatrix}.$$

We wish to compute all least squares solutions of $Ax = b$ using Theorem 7.27. Fortunately, we have already computed the SVD of A in Example 59 as $A = U\Sigma V^T$ where

$$U = \begin{bmatrix} \frac{1}{3} & -\frac{2}{\sqrt{5}} & -\frac{2}{3\sqrt{5}} \\ \frac{2}{3} & \frac{1}{\sqrt{5}} & -\frac{4}{3\sqrt{5}} \\ \frac{2}{3} & 0 & \frac{5}{3\sqrt{5}} \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sqrt{18} & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{and} \quad V = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

Thus, we merely compute solutions from (7.11), and since $k = 1$ this is merely

$$x = \frac{c_1}{\sigma_1} v_1 + y_2 v_2$$

where y_2 is any real number. In this case, we have $\sigma_1 = \sqrt{18}$ and

$$c_1 = u_1^T b = \frac{1}{3} \begin{bmatrix} 1 & 2 & 2 \end{bmatrix} \begin{bmatrix} 15 \\ 15 \\ -30 \end{bmatrix} = -5$$

so that

$$\begin{aligned} x &= \frac{-5}{\sqrt{18}} \cdot \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + \frac{y_2}{\sqrt{2}} \begin{bmatrix} 1 \\ -1 \end{bmatrix} \\ &= -\frac{5}{6} \begin{bmatrix} 1 \\ 1 \end{bmatrix} + t \begin{bmatrix} 1 \\ -1 \end{bmatrix} \end{aligned}$$

where we have replaced $\frac{1}{\sqrt{2}} y_2$ with the arbitrary parameter $t \in \mathbb{R}$. Of course, the solution which minimizes the 2-norm is

$$x_{\min} = -\frac{5}{6} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

As previously discussed in the Least Squares section, the additional least squares solutions are created by moving through the subspace

$$\text{Nul}(A) = \left\{ t \begin{bmatrix} 1 \\ -1 \end{bmatrix} : t \in \mathbb{R} \right\}$$

as can be seen from the solution representation above.

As we've likely all noticed at one point or another, many Linear Algebra students desperately want to invert non-square matrices. Of course, this cannot be done in a conventional sense, but that doesn't stop them from trying. It turns out that they're not too far off, as there is a generalized way to "invert" non-square matrices, and this is referred to as the Moore-Penrose pseudoinverse (MPP).

Definition 7.14. Given $A \in \mathbb{R}^{p \times q} \setminus \{0\}$, the **Moore-Penrose pseudoinverse** or **MPP** is defined to be the unique $A^+ \in \mathbb{R}^{q \times p}$ satisfying

$$A^+ = \underset{B \in \mathbb{R}^{q \times p}}{\arg\min} \|AB - \mathbb{I}_p\|_F$$

where $\|\cdot\|_F$ is the Frobenius norm.

Of course, it can be difficult to solve minimization problems like this, and while alternative (yet equivalent) definitions do exist, perhaps one of the easiest methods for computing A^+ is in terms of the SVD of A.

Theorem 7.28. Let $A \in \mathbb{R}^{p\times q}$ be given with SVD $A = U\Sigma V^T$ where $\Sigma \in \mathbb{R}^{p\times q}$ is defined by

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & ... & ... & 0 \\ 0 & \sigma_2 & 0 & ... & 0 \\ \vdots & 0 & \ddots & ... & 0 \\ 0 & ... & 0 & \sigma_k & 0 \\ 0 & ... & ... & ... & 0 \end{bmatrix}$$

and k is the number of nonzero singular values of A. Then, the Moore-Penrose pseudoinverse of A is

$$A^+ = V\Sigma^+ U^T$$

where $\Sigma^+ \in \mathbb{R}^{q\times p}$ is defined by

$$\Sigma^+ = \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & \dots & 0 \\ \vdots & 0 & \ddots & \dots & 0 \\ 0 & \dots & 0 & \frac{1}{\sigma_k} & 0 \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}.$$

Proof. With the above definition, we merely need to prove that

$$\|AA^+ - \mathbb{I}_p\|_F \leq \|AY - \mathbb{I}_p\|_F$$

for any $Y \in \mathbb{R}^{q\times p}$. The first step is to use the SVD of A, the construction of A^+, and the identity $UU^T = \mathbb{I}_p$ to rewrite the statement of minimization. Further, we use Lemma 7.25 with these ingredients to find

$$\begin{aligned} \|AA^+ - \mathbb{I}_p\|_F &= \|U\Sigma V^T V\Sigma^+ U^T - UU^T\|_F \\ &= \|U\Sigma\Sigma^+ U^T - UU^T\|_F \\ &= \|U\left(\Sigma\Sigma^+ - \mathbb{I}_p\right)U^T\|_F \\ &= \|\left(\Sigma\Sigma^+ - \mathbb{I}_p\right)U^T\|_F \\ &= \|\Sigma\Sigma^+ - \mathbb{I}_p\|_F. \end{aligned}$$

Next, we note that $\Sigma\Sigma^+ \in \mathbb{R}^{p\times p}$ is just

$$\Sigma\Sigma^+ = \begin{bmatrix} \sigma_1 & 0 & ... & ... & 0 \\ 0 & \sigma_2 & 0 & ... & 0 \\ \vdots & 0 & \ddots & ... & 0 \\ 0 & ... & 0 & \sigma_k & 0 \\ 0 & ... & ... & ... & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sigma_1} & 0 & \dots & \dots & 0 \\ 0 & \frac{1}{\sigma_2} & \dots & \dots & 0 \\ \vdots & 0 & \ddots & \dots & 0 \\ 0 & \dots & 0 & \frac{1}{\sigma_k} & 0 \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ 0 & \ddots & \dots & \dots & 0 \\ \vdots & 0 & 1 & \dots & 0 \\ \vdots & \dots & \dots & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

which is exactly $\mathbb{I}_p$ with the last $p-k$ diagonal entries set to zero. Hence, subtracting the identity from this results in a matrix with $p-k$ entries possessing

the value -1 along the diagonal and only entries of zeros elsewhere. The Frobenius norm of this matrix, which is exactly the square root of the sum of the squares of all entries, must then be $\sqrt{p-k}$. Therefore, we conclude

$$\|AA^+ - \mathbb{I}_p\|_F = \sqrt{p-k}.$$

Next, we show that $\sqrt{p-k}$ is actually a lower bound for $\|AY - \mathbb{I}_p\|_F$. Let $Y \in \mathbb{R}^{q\times p}$ be given and define $B \in \mathbb{R}^{q\times p}$ by $B = V^TYU$. Using similar tools as before, we compute

$$\begin{aligned}
\|AY - \mathbb{I}_p\|_F &= \|U\Sigma V^TY - UU^T\|_F \\
&= \|U\left(\Sigma V^TYU - \mathbb{I}_p\right)U^T\|_F \\
&= \|\left(\Sigma V^TYU - \mathbb{I}_p\right)U^T\|_F \\
&= \|\Sigma V^TYU - \mathbb{I}_p\|_F \\
&= \|\Sigma B - \mathbb{I}_p\|_F.
\end{aligned}$$

Then, because rows $k+1$ through p of Σ are all rows of zeros, the product $\Sigma B \in \mathbb{R}^{p\times p}$ must also have rows $k+1$ through p as rows of zeros. Therefore, these same rows in the matrix $\Sigma B - \mathbb{I}_p$ must have the value -1 along their diagonal entries. Hence, denoting the entries of $\Sigma B - \mathbb{I}_p$ by q_{ij}, we find

$$\|\Sigma B - \mathbb{I}_p\|_F = \sqrt{\sum_{i=1}^{p}\sum_{j=1}^{p}|q_{ij}|^2} = \sqrt{\sum_{i=1}^{k}\sum_{j=1}^{p}|q_{ij}|^2 + (p-k)} \geq \sqrt{p-k}.$$

Combining this inequality with the previous one, we have

$$\|AA^+ - \mathbb{I}_p\|_F = \sqrt{p-k} \leq \|AY - \mathbb{I}_p\|_F$$

for any $Y \in \mathbb{R}^{q\times p}$, and the proof is complete. □

Hence, the SVD provides an easy solution to finding the MPP. As you might further guess, the pseudoinverse solution of the inconsistent linear system $Ax = b$ is also intimately related to least-squares solutions of this problem.

In addition to each of these properties involving the SVD, there is another item of interest, referred to as Principle Components Analysis (PCA), that will be discussed in greater detail within the final chapter. More specifically, we will explore one particular application of PCA to the field of image compression.

Exercises - Operator Decompositions and Factorizations

Problem 7.1. Recall that for $A \in \mathbb{C}^{p\times p}$, the spectrum of A is

$$\sigma(A) := \Big\{\lambda \in \mathbb{C} : Ax = \lambda x \text{ for some } x \in \mathbb{C}^p \setminus \{0\}\Big\}.$$

(a) Show that $\sigma(A^T) = \sigma(A)$ for any $A \in \mathbb{C}^{p\times p}$.
Hint: $\det(C^T) = \det(C)$ for any $C \in \mathbb{C}^{p\times p}$.

(b) Assume $B \in \mathbb{C}^{p\times p}$ is nonsingular. Show that if $\lambda \in \sigma(B)$ then $\lambda^{-1} \in \sigma(B^{-1})$.

Problem 7.2. Let $A \in \mathbb{R}^{q\times q}$ be given with eigenvalue $\lambda \in \mathbb{C}$. Show that

$$|\lambda| \leq \|A\|_p$$

for every $p \in [1, \infty]$.

Problem 7.3. Prove Lemma 7.2. In particular, let $T : \mathcal{V} \to \mathcal{V}$ be linear and $k \in \mathbb{N}$. Show that if $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of T and $v_1, \ldots, v_k$ are any associated eigenvectors, then the set $S = \{v_1, \ldots, v_k\}$ is linearly independent.

Problem 7.4. Recall that $A \in \mathbb{C}^{p\times p}$ is diagonalizable if a basis for $\mathbb{C}^p$ can be formed from a collection of the eigenvectors of A. Prove that $A \in \mathbb{C}^{p\times p}$ is diagonalizable if and only if there exists a nonsingular $P \in \mathbb{C}^{p\times p}$ and diagonal $\Lambda \in \mathbb{C}^{p\times p}$ such that

$$A = P\Lambda P^{-1}.$$

Problem 7.5. Let $P \in \mathbb{R}^{p\times p}$ be given. Prove that P is orthogonal if and only if its columns are orthonormal with respect to the standard inner product (and associated norm) on $\mathbb{R}^p$.

Problem 7.6. Assume $A \in \mathbb{C}^{p\times p}$ is a unitary matrix, and let $\langle x, y\rangle$ and $\|x\|_2$ for $x, y \in \mathbb{C}^p$ represent the standard inner product and norm on $\mathbb{C}^p$, respectively. Additionally, let $\|A\|_2$ represent the associated operator (matrix) norm for any $A \in \mathbb{C}^{p\times p}$. Prove the following results without invoking analogous theorems presented in class about unitary operators:

(a) For all $x, y \in \mathbb{C}^p$

$$\langle Ax, Ay\rangle = \langle x, y\rangle$$

(b) For all $x \in \mathbb{C}^p$

$$\|Ax\|_2 = \|x\|_2$$

(c) $\|A\|_2 = 1$

(d) If $\lambda \in \mathbb{C}$ is any eigenvalue of A, then $|\lambda| = 1$.

Problem 7.7. (a) Prove that if $A, B \in \mathbb{C}^{p\times p}$ are unitary, then AB is unitary.

(b) Find $z_1, z_2 \in \mathbb{C}$ such that U is unitary where

$$U = \begin{bmatrix} \frac{1}{\sqrt{7}}(1+2i) & z_1 \\ \frac{1}{\sqrt{7}}(1-i) & z_2 \end{bmatrix}.$$

Problem 7.8. Use the method presented in class to find a Schur form of the matrix

$$A = \begin{bmatrix} 2 & 2 & -6 \\ 2 & -1 & -3 \\ -2 & -1 & 1 \end{bmatrix}.$$

To make the calculations easier, first verify that $\lambda = -2$ and $x = [1, -2, 0]^T$ are an eigenvalue/eigenvector pair for A by using the definition. For the sake of your sanity, I will mention that you should eventually find T of the form

$$T = \begin{bmatrix} -2 & 0 & 0 \\ 0 & -2 & \gamma \\ 0 & 0 & \delta \end{bmatrix}.$$

Problem 7.9. Define $A \in \mathbb{C}^{2\times 2}$ by

$$A = \begin{bmatrix} 2 & -2i \\ 2i & 2 \end{bmatrix}.$$

Show that A is normal and construct a unitary diagonalization of A; that is, find a diagonal matrix D and unitary matrix P such that $A = PDP^H$.

Problem 7.10. Let $\mathcal{V}$ be a complex Hilbert space.

(a) Show that any normal operator $T : \mathcal{V} \to \mathcal{V}$ satisfies $\text{Ker}(T) = \text{Ker}(T^*)$.
Hint: Use a lemma from class concerning $\|T(v)\|$.

(b) Let $T : \mathcal{V} \to \mathcal{V}$ be normal. Show that if v is an eigenvector of T with corresponding eigenvalue λ, then v is an eigenvector of T^* with corresponding eigenvalue $\bar{\lambda}$.
Hint: Use the result of part (a).

(c) Let $\lambda_1 \neq \lambda_2$ be eigenvalues of the normal operator $T : \mathcal{V} \to \mathcal{V}$ with corresponding eigenvectors $v_1, v_2 \in \mathcal{V}$, respectively. Show that v_1 and v_2 are orthogonal.
Hint: Use the result of part (b) to show $(\lambda_1 - \lambda_2)\langle v_1, v_2\rangle = 0$.

Problem 7.11. Let $\mathcal{V}$ be a Hilbert space and $T : \mathcal{V} \to \mathcal{V}$ be a normal, bounded linear operator. Show that for every $m \in \mathbb{N}$

$$\|T^{2^m}\|^2 = \|(T^{2^m})^* T^{2^m}\|.$$

Problem 7.12. Let $\mathcal{V}$ be a Hilbert space and $T : \mathcal{V} \to \mathcal{V}$ be a bounded linear operator defined by

$$T(v) = \sum_{n=1}^{\infty} \lambda_n \langle e_n, v\rangle e_n$$

for every $v \in \mathcal{V}$, where $e_n \in H$ is an orthonormal sequence of eigenvectors of T with corresponding eigenvalues $\lambda_n \in \mathbb{C}$. Show that

$$T(T^*(v)) = \sum_{n=1}^{\infty} |\lambda_n|^2 \langle e_n, v\rangle e_n$$

for every $v \in \mathcal{V}$. You may assume the formula for T^* is known from the proof of Theorem 7.17.

Problem 7.13. Let $u \in \mathbb{R}^p$ and $v \in \mathbb{R}^q$ be unit vectors. Show that $A \in \mathbb{R}^{p\times q}$ defined by $A = u \otimes v = uv^T$ satisfies

$$\|A\|_2 = 1.$$

Problem 7.14. Construct the matrices U, Σ, V in the singular value decomposition $A = U\Sigma V^T$ of

$$A = \begin{bmatrix} 4 & -2 \\ 2 & -1 \\ 0 & 0 \end{bmatrix}.$$

Problem 7.15. Let $A \in \mathbb{R}^{2\times 100}$ be defined by

$$\begin{cases} A_{1k} = k, & \text{for } k = 1, ..., 100 \\ A_{2k} = 2k, & \text{for } k = 1, ..., 100. \end{cases}$$

(a) How many real scalars are needed to store A?

(b) Compute an SVD of A. It may be helpful to use the outer product notation for this decomposition.

(c) How many real scalars are needed to store the SVD of A?

Problem 7.16. Let $\lambda \in \mathbb{C}$ and $v \in \mathbb{C}^p \setminus \{0\}$ be any eigenpair of the normal matrix $B \in \mathbb{C}^{p\times p}$. Prove that $|\lambda|$ is a singular value of B.

Problem 7.17. Let $A \in \mathbb{R}^{p\times p}$ be given and $\lambda \in \mathbb{C}$ be any eigenvalue of A. Use a Singular Value Decomposition to prove

$$|\lambda| \leq \sigma_1$$

where σ_1 is the first singular value of A.

Problem 7.18. Let $A \in \mathbb{C}^{p\times q}$ be given and use an SVD to prove that

$$\|A\|_2 = \|A^H\|_2.$$

Problem 7.19. Let $A \in \mathbb{R}^{p\times q}$ be given. Show that

$$\|A\|_F \leq \sqrt{\text{rank}(A)}\|A\|_2.$$

Problem 7.20. Let $A \in \mathbb{R}^{p\times p}$ be given with singular value decomposition $A = U\Sigma V^T$. Show that $Q_A = UV^T$ is orthogonal and that

$$\|Q_A - A\|_F \leq \|Q - A\|_F$$

for every orthogonal $Q \in \mathbb{R}^{p\times p}$. In this way, Q_A is the best orthogonal approximation to A in the Frobenius norm.

Problem 7.21. Let $A \in \mathbb{C}^{p\times q}$ with $\text{rank}(A) = q$.

(a) Show that A^HA is nonsingular and that $A^+ = (A^HA)^{-1}A^H$.

(b) Assume that $A = QR$ where $Q \in \mathbb{C}^{p\times q}$ has orthonormal columns (with respect to the standard inner product on $\mathbb{C}^p$) and $R \in \mathbb{C}^{q\times q}$. Show that R is nonsingular and $A^+ = R^{-1}Q^H$.

Chapter 8

Application: Principal Component Analysis

In the previous chapter, we were introduced to the Singular Value Decomposition (SVD) within Theorem 7.20. This fundamental result (and the comments following its proof) demonstrates that any real matrix can be decomposed into the linear combination of outer products of eigenvectors of its Gram matrices (i.e. A^TA and AA^T), where the singular values of the given matrix are exactly the coefficients in this linear combination. Said another way, any matrix A can be written as

$$A = \sum_{j=1}^{k} \sigma_j u_j \otimes v_j \tag{8.1}$$

or equivalently

$$A = \sigma_1 u_1 v_1^T + \sigma_2 u_2 v_2^T + ... + \sigma_k u_k v_k^T$$

where σ_j is the jth of k singular values of A, and the u_j and v_j vectors are corresponding eigenvectors of AA^T and A^TA, respectively. Note that $\|u_j \otimes v_j\|_2 = 1$ for every $j = 1, ..., k$. Additionally, the singular values are placed into decreasing order, and hence the importance of the information within A is arranged in the same manner. Often, we can compress the storage of a matrix $A \in \mathbb{R}^{p\times q}$ with minimal loss of information by truncating this decomposed sum (8.1). Such a technique is sometimes referred to as Principal Components Analysis (PCA) and has been around since the turn of the nineteenth century (invented by Karl Pearson in 1901 [30]).

PCA arises in countless disciplines, including but not limited to statistics [6], electrical engineering and control theory [25], genetics [20], neuroscience [19], facial recognition [28], and mechanical and systems engineering [4], among others. In these fields, the method is often used to compress data or images, process signals, conduct remote sensing activities, or perform statistical analyses. Principal Components Analysis is so widely utilized, in fact, that it has been given a variety of monikers within these different fields, including the Proper Orthogonal Decomposition (POD), discrete Karhunen-Loéve transform (KLT), Hotelling transform, Empirical Orthogonal Function (EOF) Analysis, empirical eigenfunction decomposition, empirical component analysis, Eckart-Young theorem (truncated SVD), and methods of "factor analysis". For reasons that we will see later, it is often also mistaken for the SVD because the two are so inherently tied.

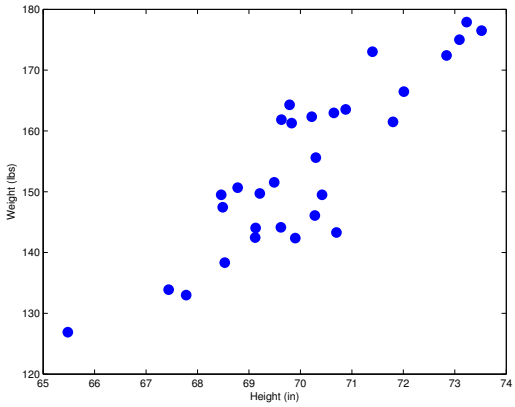


Figure 8.1: Plot of Height/Weight datapoints.

Additional information and applications concerning PCA can be found in the review article [18].

Of course, PCA can also be thought of not as a transformation of the given matrix A, but in terms of the Gram (or covariance) matrix $C = AA^T$ (alternatively, the matrix A^TA could be used). In this case, the fact that $C \in \mathbb{R}^{p\times p}$ is necessarily symmetric can be exploited, and in view of Corollary 7.4, the symmetry property is equivalent to C being orthogonally diagonalizable. This is an extremely powerful result and precisely guarantees, for such C, the existence of $\lambda_j \in \mathbb{R}$ and orthonormal column vectors $v_j \in \mathbb{R}^p$ for every $j = 1, ..., p$ such that

$$C = \sum_{j=1}^{p} \lambda_j v_j v_j^T = \lambda_1 v_1 v_1^T + \cdots + \lambda_n v_p v_p^T \tag{8.2}$$

arising from the orthogonal diagonalization. Because the v_j vectors are orthonormal, each associated matrix $v_j v_j^T$ is orthogonal, and thus C can be decomposed into a sum wherein each term is an eigenvalue multiplied by a rank one matrix generated by a unit vector. Hence, the eigenvalues alone determine the magnitude of each term in the sum, while the eigenvectors determine the directions. These eigenvectors are called *principal components* or *principal directions* and we will expand upon this further in the next section. We begin the discussion of the specifics of PCA with an introductory example and later return to the interpretation of this method that involves the SVD.

8.1 Introductory Height & Weight Problem

Consider a study in which we want to determine whether or not the heights and weights of a group of individuals are correlated. That is, we want to know whether the known value of a person's height seems to dictate whether they tend to be heavier or lighter, and thus influences their weight. Assume we are given data for 30 specific people, displayed within Table 8.1. For this example, our data set originates from a commonly available study [22], though we could just as easily collect it from our class.

Since the question of interest is whether the two measured variables, height and weight, seem to change together, the relevant quantity to consider is the covariance

Person	1	2	3	4	5	6
Height	67.78	73.52	71.40	70.22	69.79	70.70
Weight	132.99	176.49	173.03	162.34	164.30	143.30
Person	7	8	9	10	11	12
Height	71.80	72.01	69.90	68.78	68.49	69.62
Weight	161.49	166.46	142.37	150.67	147.45	144.14
Person	13	14	15	16	17	18
Height	70.30	69.12	70.28	73.09	68.46	70.65
Weight	155.61	142.46	146.09	175.00	149.50	162.97
Person	19	20	21	22	23	24
Height	73.23	69.13	69.83	70.88	65.48	70.42
Weight	177.90	144.04	161.28	163.54	126.90	149.50
Person	25	26	27	28	29	30
Height	69.63	69.21	72.84	69.49	68.53	67.44
Weight	161.85	149.72	172.42	151.55	138.33	133.89

Table 8.1: Heights (in.) and Weights (lbs.) for 30 young adults [22].

of the two characteristics within the data set. This can be formed in the following way. First, the data is stored in a 2×30 matrix A. Then, the entries are used to compute the mean in each row, which will be used to center or "mean-subtract" the data. This latter step is essential, as many of the results concerning PCA are only valid upon centering the data at the origin. Computing the means of our measurements (Table 8.1), we find

$$\mu = \begin{bmatrix} 70.06 \\ 154.25 \end{bmatrix}.$$

Using a_{ij}, the entries of the data matrix A, the associated 2×2 covariance matrix S is constructed with entries

$$s_{ik} = \frac{1}{30-1} \sum_{j=1}^{30} (a_{ij} - \mu_i)(a_{kj} - \mu_k)$$

so that

$$S = \begin{bmatrix} 3.26 & 21.72 \\ 21.72 & 188.96 \end{bmatrix}.$$

Notice that this matrix is necessarily symmetric, so using the Spectral Theorem it can be orthogonally diagonalized. Upon computing the eigenvalues and eigenvectors of S, we find

$$\lambda_1 = 191.46 \qquad \text{and} \qquad \lambda_2 = 0.76,$$

and

$$v_1 = \begin{bmatrix} 0.11 \\ 0.99 \end{bmatrix} \qquad v_2 = \begin{bmatrix} -0.99 \\ 0.11 \end{bmatrix}.$$

Here, v_1 and v_2 are the *principal components* of the covariance matrix S generated by the data matrix A, as previously described. Obviously, these vectors form an orthogonal set. Thus, we see from (8.2) that

$$S = \lambda_1 v_1 v_1^T + \lambda_2 v_2 v_2^T$$

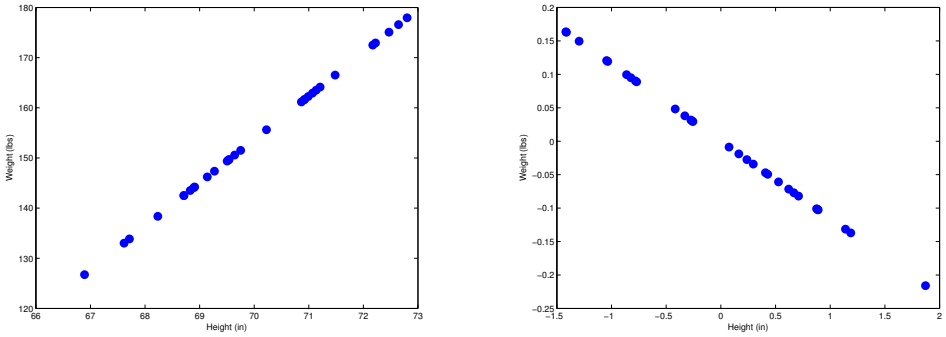


Figure 8.2: Height/Weight data projected onto the principal components (left - v_1; right - v_2). By the Spectral Theorem, the data represented in Fig. 8.1 is exactly the sum of the projections onto these two components.

and because the difference in eigenvalues is so large (i.e. $\lambda_1 \gg \lambda_2$), it appears that the first term is responsible for most of the information encapsulated within S. Regardless, we can re-express the given data in the new orthonormal basis generated by v_1 and v_2 by computing the coordinates $P^T A$ where

$$P = \begin{bmatrix} 0.11 & -0.99 \\ 0.99 & 0.11 \end{bmatrix}$$

is the orthogonal matrix whose columns are v_1 and v_2. In fact, we could left multiply the data matrix by each component separately, namely $v_1^T A$ and $v_2^T A$, to project the data onto each principal direction (Fig. 8.2). Hence, the data can be separated into projections along v_1 and v_2, respectively. We see from looking at the scales in Figure 8.2 that the heights and weights along v_2 are significantly less than those along v_1, which tells us that the majority of the information contained within A lies along v_1. Computing the slope of the line in the direction of v_1 and choosing a point through which it passes, we can represent the equation of the line by

$$y - 154.25 = 9(x - 70.06),$$

where x represents the height of a given individual and y is their corresponding weight. Hence, we see that height and weight appear to be strongly correlated, and PCA has determined the direction with optimal correlation (generated by our sample) between the variables.

The principal component analysis for this example took a small set of data and identified a new orthonormal basis in which to re-express it. In two dimensions the data are effectively rotated to lie along the line of best fit (Fig. 8.3), with the second principal direction merely representing the associated unit orthogonal complement of the first. This mirrors one general aim of PCA: to obtain a new orthonormal basis that organizes the data optimally, in the sense that the variance contained within the vectors is maximized along successive principal component(s).

8.2 Summary of PCA

In short, PCA can be performed to compute an optimal, ordered orthonormal basis of a given set of vectors, or data set, in the following steps.

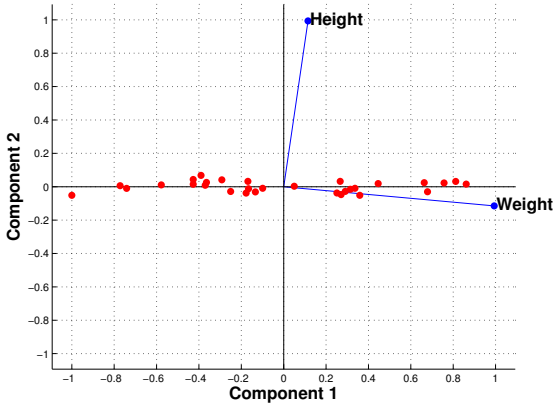


Figure 8.3: Biplot of Height/Weight data with 2 principal components. The blue Height and Weight vectors are displayed as linear combinations of the principal components. Note that these principal components effectively rotate the height and weight data in the plane.

1. Gather n samples of m-dimensional data, i.e. vectors $d_1, ..., d_n \in \mathbb{R}^m$ stored in the $m \times n$ matrix A with columns $d_1, ..., d_n$, so that a_{ij} represents the i^{th} entry of the j^{th} sample vector, and compute the mean vector (in $\mathbb{R}^m$)

$$\mu = \frac{1}{n} \sum_{k=1}^{n} d_k,$$

2. Build the corresponding mean-centered data matrix B with columns given by $d_j - \mu$ so that the entries are

$$b_{ij} = x_{ij} - \mu_i$$

for every $i = 1, ..., m$ and $j = 1, ..., n$.

3. Use B to compute the symmetric, $m \times m$ covariance matrix

$$S = \frac{1}{n-1} BB^T.$$

4. Find the eigenvalues $\lambda_1, ..., \lambda_m$ of S (arranged in decreasing order including multiplicity) and an orthonormal set of corresponding eigenvectors $v_1, ..., v_m$. These create a new basis for $\mathbb{R}^m$ in which the data can be expressed.

5. Finally, the data is represented in terms of the new basis vectors $v_1, ..., v_m$ using the coordinates $y_1 = v_1^T A, ..., y_m = v_m^T A$. This can also be represented as the matrix $Y = P^T A$ where P is the matrix with columns $v_1, ..., v_m$. Should we wish to convert the data back to the original basis, we merely utilize the orthogonality of P and compute $PY = PP^T Y = A$ to find the original data matrix A.

In the final section, we will extend our introductory example while presenting an additional application in which PCA appears prominently and can be easily visualized, namely image compression.

8.3 PCA for Image Compression and the SVD

Another important application of PCA is Image Compression. Because images are stored as large matrices with real entries, one can reduce their storage requirements by retaining only the essential portions of the image [11]. Of course, information (in this case, fine-grained detail of the image) is naturally lost in this process, but it is done in an optimal manner, so as to preserve the most essential characteristics of the original image. In this section we detail a specific example for the use of PCA to compress an image. Since the effects of keeping a lower dimensional projection of the image will be visually clear, this particular example is a great in-class activity.

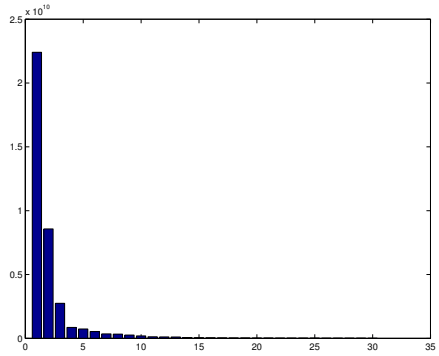
Throughout the example we will work with a built-in test image - Albrecht Dürer's *Melancholia* displayed in Fig. 8.4a. MATLAB considers images like this as objects consisting of two portions - a matrix of pixels and a colormap. Our image is stored in a 648×509 pixel matrix, and thus contains $648 \times 509 = 329,832$ total pixels. The colormap is a 648×3 matrix, which we will ignore for the current study. Each element of the pixel matrix contains a real number representing the intensity of grey scale for the corresponding pixel. MATLAB displays all of the pixels simultaneously with the correct intensity, and the greyscale image that we see is produced. The 648×509 matrix containing the pixel information is our data matrix, A, which unlike the previous section will not require centering.

To compress the image, we wish to construct a reduced representation of A, namely some matrix $\tilde{A}$ that requires us to store fewer values. Of course, PCA provides us with a way to determine and capture the most important information in A and then store this in the reduced matrix $\tilde{A}$ using the first few principal components. In this way, the original image stored by A will be compressed by using the lower-rank approximation $\tilde{A}$.

In the previous section, we developed a method for PCA which uses the given data matrix to determine a new orthonormal basis that captures maximal variance. Since the associated eigenvalues are listed in decreasing order, we might also be able to truncate the sum in (8.2) to reduce the amount of stored data. For instance, in the height/weight example, the first principal component contained the overwhelming majority of the information embedded within the data. Hence, we might only keep this vector v_1 and discard v_2 since the data can be mostly explained just by knowing the former characteristic, rather than every height and weight. In this case, each data point would then be represented by its projection onto the first principal component. Upon performing this step, we might also interpret the results: are a small number of the eigenvalues λ_j much less (perhaps by an order of magnitude) than the others? If so, this indicates that a reduction in the dimension of the data is possible without losing too much information, while if this does not occur then the dimension of the data may not be easily reduced in such a way.

Of course, it can also be tedious to even compute the covariance matrix and its eigenvalues in the first place. As previously mentioned, the Dürer image is represented by a 648×509 matrix. If we denote this by A, then it will require some work to both generate the matrix $A^T A \in \mathbb{R}^{509 \times 509}$ and compute its eigenvalues.

(a) Dürer's *Melancholia*



(b) Eigenvalues

Figure 8.4: (a) Albrecht Dürer's *Melancholia* displayed as a 648×509 pixelated image, taken from Matlab's built-in "Durer" file. (b) The first 35 eigenvalues of the covariance matrix S generated from A in the image compression example and arranged in decreasing order.

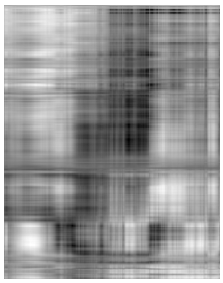
From the previous chapter, however, we know that these eigenvalues are merely the squares of the singular values of A. So, if accurate and efficient computational methods exist to compute the SVD (and they do), then it would likely be easier to obtain the singular values without even constructing a covariance matrix. Additionally, the left and right singular vectors of A are really just eigenvectors of AA^T and A^TA, respectively, which means that these can also be obtained without forming these larger matrices. Hence, using the SVD to generate this information will likely be easier and faster than the method described in the previous section.

In this vein, let's return to (8.1) and suppose that instead of computing all of the σ_j singular values and u_j and v_j vectors in the sum, we merely choose the first ℓ of each of these with $\ell \ll k$. Using these singular values and principal components, we could then construct a truncated representation of A, denoted by $\tilde{A}$, but only including the most dominant ℓ terms in the sum. Stated another way, if $A = U\Sigma V^T$ is the SVD of A with $U \in \mathbb{R}^{p\times k}$, $\Sigma \in \mathbb{R}^{k\times k}$, and $V \in \mathbb{R}^{q\times k}$, then we have constructed $\tilde{U} \in \mathbb{R}^{p\times \ell}$, $\tilde{\Sigma} \in \mathbb{R}^{\ell\times \ell}$, and $\tilde{V} \in \mathbb{R}^{q\times \ell}$ such that

$$\tilde{A} = \tilde{U}\tilde{\Sigma}\tilde{V}^T.$$

Here, the retained u_j and v_j vectors form the ℓ columns of $\tilde{U}$ and $\tilde{V}$, respectively, while the σ_j values form the diagonal of $\tilde{\Sigma}$. In this way, we keep only the first ℓ columns of the original U and V orthogonal matrices, as well as, the first ℓ columns and rows of the original Σ with $\ell \ll k$. Notice that $\tilde{A}$ is still a 648×509 matrix so that the dimensions of the original image have not been changed. However, to reconstruct $\tilde{A}$ we need only know the ℓ columns of U and V and ℓ singular values, whereas reconstructing A requires k of these. In terms of actual storage $\tilde{A}$ requires knowledge of $\ell(1 + p + q)$ numbers and A requires knowledge of $k(1 + p + q)$ numbers. Hence, the amount of information needed to capture $\tilde{A}$ is drastically reduced, and the larger p and q are or the larger the difference between ℓ and k, the more compressed the approximation becomes.

With a method for determining an efficient approximation, the next question should likely address how accurate or precise this truncation will be. In fact, an



(a) $\ell = 3$, $\sigma^2 = 88.89\%$

(b) $\ell = 30$, $\sigma^2 = 99.61\%$

(c) $\ell = 60$, $\sigma^2 = 99.78\%$

(d) $\ell = 90$, $\sigma^2 = 99.92\%$

Figure 8.5: The Dürer image with varying numbers of principal components.

error estimate is obtained from the SVD as well - namely, the amount of information retained is given by the *spectral ratio* of the associated covariance matrix, defined by

$$\sigma^2 := \frac{\sum_{j=1}^{\ell} \lambda_j}{\sum_{j=1}^{k} \lambda_j}$$

Furthermore, this quantity can be better represented in terms of the singular values of A by

$$\sigma^2 = \frac{\sum_{j=1}^{\ell} \sigma_j^2}{\sum_{j=1}^{k} \sigma_j^2}. \tag{8.3}$$

Thus, in our height/weight example, we can keep only the first component of each data point (a 1×30 matrix) rather than the full data set (2×30 matrix) and still retain 99% of the information contained within because

$$\sigma^2 = \frac{191.46}{191.46 + 0.76} > 0.99.$$

In situations where the dimensions of the data are large, but the components are highly correlated, it is beneficial to reduce the dimension of the data matrix using PCA. This has three effects: it (1) orthogonalizes the basis vectors (so that they are uncorrelated), (2) orders the resulting orthogonal components so that those with the largest variance appear first, and (3) eliminates dimensions that contribute the least to the variation in the original data set.

To perform this method on the Dürer image, we use the SVD to compute the singular values of A and use these to form the distribution of eigenvalues of the covariance matrix $S = A^T A$. As displayed in Fig. 8.4b doing so shows us the

formation of a large *spectral gap*, i.e. a large difference between consecutive eigenvalues. Therefore, if we write the SVD of A as in (8.1), namely

$$A = \sum_{j=1}^{k} \sigma_j u_j \otimes v_j$$

and then create the reduced approximation

$$\tilde{A} = \sum_{j=1}^{\ell} \sigma_j u_j \otimes v_j$$

by truncating this sum at a particular index ℓ with $\ell \ll k$, the resulting linear combination of principal components in $\tilde{A}$ will still contain a large amount of the total information embedded within the original image A. This occurs because the terms we have eliminated in the representation of A will be scaled by the coefficients $\sigma_{\ell+1}, ..., \sigma_k$ which are significantly smaller than the σ_j terms that we've kept within the truncated representation of $\tilde{A}$. The MATLAB code to compute the spectral ratio and display differing reduced-rank image approximations to A is given below.

In Fig. 8.5, we've represented $\tilde{A}$ for four different choices of ℓ (i.e., the number of principal components used), and the associated spectral ratio, σ^2, retained by those reduced descriptions is also listed. Notice that the detail of the image improves as ℓ is increased, and that a fairly suitable representation can be obtained with around 90 components rather than the full 648 vector description. Thus, PCA has served the useful purpose of reducing the dimension of the original data set while preserving its most essential features.

```
clear; clc;
load durer
size(A)

totNumEntries = size(A,1)*size(A,2)

fig1=figure('Color',[1 1 1]);
image(A), colormap(map), axis off, axis equal;

%A = double(A(:, :, 2));
A = double(A);

%
[U,S,V] = svd(A);
sing = diag(S);
fig2=figure;
plot(log10(sing)),...
    ylabel('log$_{10}(\sigma_j)$', 'interpreter', 'latex', ...
        'FontSize',14),...
    xlabel('$j$','interpreter','latex', 'FontSize',14)

maxSing = sing(1)
minSing = sing(end)
sumSing = cumsum(sing);

% Find number of singluar values to retain "90%" of image
trunc = min(find(sumSing > 0.9*sumSing(end)))

k=246;
totvar = sum(sing);
percentkept = sum(sing(1:k))/totvar;
entriesStored = k*(size(A,1)+size(A,2)+1);
compressRatio = entriesStored/totNumEntries;

B = U(:,1:k)*S(1:k,1:k)*V(:,1:k)';
R = double(A)-double(B);
relL2Error = norm(R,2)/norm(A,2);

fig3=figure('Color',[1 1 1]);
B = uint8(B);
image(B), colormap(map), axis off, axis equal
title({['Singular values used = ', num2str(100*percentkept), ...
    '%'], ...
    ['Compression Ratio = ', num2str(100*compressRatio), '%'], ...
    ['Relative L2 Error = ', num2str(100*relL2Error), '%']});

R = uint8(R);
fig4=figure('Color',[1 1 1]);
image(R), colormap(map), axis off, axis equal
title(['Difference, k= ', num2str(k)]);
```

Chapter 9

Appendix

In the appendix, we state some informative definitions and celebrated results from Real & Complex Analysis. Throughout, we let $n \in \mathbb{N}$ and take $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ as in the main text.

Definition 9.1. Let $A \subset \mathbb{R}$ be given.

1. We define (if it exists) the **supremum** of A, written

$$\sup\{x : x \in A\} \quad \text{or} \quad \sup_{x \in A} x,$$

to be an upper bound of A such that for any upper bound $u \in \mathbb{R}$, we have

$$\sup\{x : x \in A\} \leq u.$$

Said another way, the supremum is the least of all upper bounds of the set A.

2. Given a set $A \subset \mathbb{R}$, we define (if it exists) the **infimum** of A, written

$$\inf\{x : x \in A\} \quad \text{or} \quad \inf_{x \in A} x,$$

to be a lower bound of A such that for any lower bound $\ell \in \mathbb{R}$, we have

$$\inf\{x : x \in A\} \geq \ell.$$

Said another way, the infimum is the greatest of all lower bounds of the set A.

The following definitions provide some properties from Measure Theory.

Definition 9.2. Let a set A and function $f : A \to \mathbb{R}$ be given.

1. We define the **essential supremum** of f on A by

$$\operatorname*{ess\,sup}_{x \in A} f(x) = \inf\left\{C \in \mathbb{R} : f(x) \leq C \text{ for a.e. } x \in A\right\}.$$

2. We define the **essential infimum** of f on A by

$$\operatorname*{ess\,inf}_{x \in A} f(x) = \sup\left\{C \in \mathbb{R} : f(x) \geq C \text{ for a.e. } x \in A\right\}.$$

Definition 9.3. A set $A \subseteq \mathbb{K}^n$ is **bounded** if there exists $C > 0$ such that

$$\sum_{j=1}^{n} |x_j|^2 \leq C$$

for every vector $x \in A$.

Theorem 9.1 (Bolzano-Weierstrass)**.** Every bounded sequence $\{v_k\}_{k=1}^{\infty} \subset \mathbb{K}^n$ has a convergent subsequence.

Definition 9.4. A set $A \subseteq \mathbb{K}^n$ is **closed** if the limit of every convergent sequence $\{v_k\}_{k=1}^{\infty} \subset A$ with $v_n \to v$ in V satisfies $v \in A$. Said another way, A is closed if it contains the limits of all of its convergent sequences.

Definition 9.5. A set $A \subseteq \mathbb{K}^n$ is **compact** if every bounded sequence of A has a convergent subsequence (i.e., possesses the Bolzano-Weierstrass property).

Theorem 9.2 (Heine-Borel)**.** Let $A \subset \mathbb{K}^n$ be given. Then, A is compact if and only if A is closed and bounded.

Theorem 9.3 (Extreme Value Theorem)**.** Let A be a nonempty, compact subset of $\mathbb{K}^n$. If $f : A \to \mathbb{R}$ is continuous, then f is bounded both above and below and attains its supremum and infimum on A. Namely, there exists $a_{\min}, a_{\max} \in A$ such that

$$\sup_{x \in A} f(x) = a_{\max} \qquad \text{and} \qquad \inf_{x \in A} f(x) = a_{\min}.$$

Theorem 9.4 (Completeness of $\mathbb{K}^n$)**.** Every Cauchy sequence $\{v_k\}_{k=1}^{\infty} \subset \mathbb{K}^n$ converges to a limit $v \in \mathbb{K}^n$. That is, the normed space $\mathbb{K}^n$ is complete.

Bibliography

[1] S. Axler, *Linear Algebra Done Right*, 5th edition, Springer (2015)

[2] K. Bryan and T. Leise, *The $25,000,000,000 Eigenvector: The Linear Algebra behind Google*, (2006) SIAM Review 48(3): 569-581.

[3] S. Brin and L. Page, *The Anatomy of a Large-Scale Hypertextual Web Search Engine*. Seventh International World-Wide Web Conference (WWW 1998), April 14-18, 1998, Brisbane, Australia.

[4] B. Feeny and R. Kappagantu, *On the Physical Interpretation of Proper Orthogonal Modes in Vibrations*, Journal of Sound and Vibration 211(4): , 607-616, 1998.

[5] R. Horn and C. Johnson, *Matrix Analysis* (Section 3.5), Cambridge University Press, ISBN 978-0-521-38632-6. (1985)

[6] I. Jolliffe, *Principal Component Analysis*, Springer Series in Statistics, 2nd edition, Springer: 2002

[7] D. Kalman, *A Singularly Valuable Decomposition: The SVD of a Matrix*, (1996) The College Mathematics Journal, 27(1): 2-23.

[8] J. Keener, *The Perron-Frobenius Theorem and the Ranking of Football Teams*, (1993) SIAM Review 35(1): 80-93.

[9] V. Kisil, *Introduction to Functional Analysis*
http://www1.maths.leeds.ac.uk/~kisilv/courses/math3263.html

[10] E. Kreyszig, *Introductory Functional Analysis with Applications*, Wiley Classics Library, John Wiley & Sons, Inc., New York (1989).

[11] D. Lay, *Linear Algebra and Its Applications*, 4th edition, Pearson (2012).

[12] A. Langville and C. Meyer, *Google's PageRank and Beyond: The Science of Search Engine Rankings*, Princeton University Press (2006).

[13] A. Langville and C. Meyer, *Who's #1: The Science of Rating and Ranking*, Princeton University Press (2012).

[14] C. Meyer, *Matrix analysis and applied linear algebra*, SIAM (2000).

[15] C. Moler, *The World's Largest Matrix Computation*
https://www.mathworks.com/company/newsletters/articles/the-world-s-largest-matrix-computation.html

[16] B. Noble and J. Daniel, *Applied Linear Algebra*, 3rd edition, Pearson (1987).

[17] L. Page, S. Brin, R. Motwani and T. Winograd, *The PageRank Citation Ranking: Bringing Order to the Web*. (1999) Technical Report, Stanford InfoLab.

[18] S. Pankavich and R. Swanson *Principal Component Analysis: Resources for an Essential Application of Linear Algebra*, (2015) PRIMUS, 25:5, 400-420.

[19] A. Peyrache, K. Benchenane, M., Khamassi, S. Wiener, F., Battaglia, *Principal component analysis of ensemble recordings reveals cell assemblies at high temporal resolution* Journal of Computational Neuroscience, 29 (1-2): 309-325, 2010.

[20] S. Raychaudhuri, J. Stuart, and R. Altman, *Principal Component Analysis to summarize microarray experiments: Application to sporulation times series*, Pac Symp Biocomput 455–466, 2000.

[21] W. Rudin *Principles of Mathematical Analysis*, (1976) New York: McGraw Hill. (pp. 89-90).

[22] SOCR Data Human Heights and Weights
http://wiki.stat.ucla.edu/socr/index.php/...
SOCR_Data_Dinov_020108_HeightsWeights

[23] G. Strang, *The Fundamental Theorem of Linear Algebra*, (1993) The American Mathematical Monthly, 100 (9): 848-855.

[24] G. Strang, *Linear Algebra and its applications*, 4th ed., Cengage Learning (2005).

[25] *Principal Component Analysis - Engineering Applications*, edited by Parinya Sanguansat, InTech, 2012, DOI: 10.5772/2693

[26] L. Trefethen and D. Bau *Numerical Linear Algebra*, SIAM (1997).

[27] S. Treil, *Linear Algebra Done Wrong*, Lecture Notes - Brown University, https://www.math.brown.edu/~treil/papers/LADW/LADW.html

[28] M. Turk and A. Pentland, *Eigenfaces for recognition*, Journal of Cognitive Neuroscience, 3 (1): 71-86, 1991.

[29] C., Villani. *Optimal Transport, Old and New*, Springer (2008).

[30] Wikipedia: "Karl Pearson", Published May 08, 2019, Accessed May 28, 2019, https://en.wikipedia.org/wiki/Karl_Pearson

[31] I. Yanovsky, *Linear Algebra: Graduate Level Problems and Solutions*, hosted by UCLA,
https://www.math.ucla.edu/~yanovsky/handbooks/linear_algebra.pdf