Linear Algebra

General fields, Vector spaces, Projections, Eigenvalues and Singular values

Michael J. Neely

Fall 2026

An updated version of this text is here:*
`https://ee.usc.edu/stochastic-nets/docs/projections.pdf`

*Let me know if you find typos or mistakes so I can fix them

PREFACE ix
0.1 Philosophy . . . ix
0.2 Linear independence . . . ix
0.3 Fields . . . x
0.4 Gaussian elimination . . . xi
0.5 Checking answers with computer . . . xi
0.6 Notation . . . xi
0.7 Set or tuple? . . . xiii

I Scalars and matrices 1

1 AXIOMS FOR A FIELD 3
1.1 Summing in any order . . . 4
1.2 Multiplying in any order . . . 7
1.3 Uniqueness of additive and multiplicative inverses . . . 7
1.4 Example fields . . . 8
1.5 Basic results for fields . . . 14
1.6 Subtraction and division . . . 17
1.7 Modular arithmetic . . . 17
1.8 Exercises . . . 23

2 MATRIX OPERATIONS 29
2.1 Matrices . . . 29
2.2 Diagonal, upper triangular, and the identity matrix . . . 31
2.3 Matrix addition and scalar multiplication . . . 33
2.4 Matrix multiplication . . . 34
2.5 Data crunching with matrices . . . 40
2.6 Transpose of a product of matrices . . . 41
2.7 Trace of a matrix . . . 42
2.8 Structural properties of matrix multiplication . . . 43
2.9 Multiplying with the identity matrix . . . 46
2.10 Invertible matrices . . . 48

2.11 Permutation matrices . . . 54
2.12 Exercises . . . 57

II Subspaces and linear independence **79**

3 Subspaces of $\mathbb{F}^n$ 81
3.1 Linear combinations . . . 82
3.2 Subspace of $\mathbb{F}^n$ . . . 86
3.3 The four fundamental subspaces of a matrix A . . . 90
3.4 Speed with definitions . . . 92
3.5 Operations that do not change $Null(A)$ or $Row(A)$ . . . 93
3.6 Counting the vectors in a subspace . . . 95
3.7 Linear independence . . . 100
3.8 Spanning a subspace . . . 104
3.9 Basis of a subspace . . . 105
3.10 Proving linear independence: A case study . . . 110
3.11 Example: Establishing linear independence for two or three vectors . . . 113
3.12 Exercises . . . 116

4 Gaussian elimination 125
4.1 Particular solution and all solutions . . . 125
4.2 Back-substitution . . . 128
4.3 Overview of Gaussian elimination . . . 133
4.4 Echelon form . . . 135
4.5 Elementary row operations . . . 142
4.6 How an A becomes an R . . . 147
4.7 Examples . . . 147
4.8 Computing A^{-1} . . . 152
4.9 Bases for the fundamental subspaces of A . . . 156
4.10 Solving $AY = B$ . . . 161
4.11 Convenient factorizations . . . 163
4.12 Exercises . . . 168

5 Dimension and rank 177
5.1 Conditions for linear dependence and independence . . . 177
5.2 Wide matrix theorem . . . 179
5.3 Basis and dimension . . . 181
5.4 Growing a basis via Lemma 5.2 . . . 182
5.5 Shrinking down to a basis via Lemma 5.1 . . . 183
5.6 Using dimension to establish that vectors form a basis . . . 183

5.7 The rank of a matrix . . . 184
5.8 Equivalent conditions for invertibility . . . 187
5.9 Random matrices . . . 189
5.10 The chain of null spaces for powers of square matrices . . . 194
5.11 Exercises . . . 196

6 General vector spaces 211
6.1 Axioms for a vector space over a field . . . 211
6.2 Linear combinations . . . 213
6.3 Subspaces . . . 214
6.4 Example vector spaces . . . 215
6.5 Finite and infinite dimensional subspaces . . . 219
6.6 Linear independence, basis, and dimension . . . 220
6.7 Using dimension to establish a basis . . . 223
6.8 Examples . . . 223
6.9 Linear transformations . . . 232
6.10 $Null(T)$ and $Span(T)$ . . . 234
6.11 Example linear transformations . . . 235
6.12 Linear transformations are defined by their values on a basis . . . 238
6.13 Invertible linear transformations . . . 239
6.14 The matrix of a linear transformation . . . 239
6.15 Rank-nullity theorem for linear transformations . . . 242
6.16 Direct sums . . . 243
6.17 Dual space . . . 246
6.18 Dimension summary . . . 246
6.19 Exercises . . . 247

7 Playing with transposes and the function $x^T A y$ 261
7.1 The function $x^T A y$ . . . 263
7.2 Positive definite and positive semidefinite . . . 264
7.3 Exercises . . . 265

III Projections 269

8 Real inner product space 271
8.1 Axioms for a real inner product . . . 271
8.2 Cauchy-Schwarz inequality . . . 274
8.3 Inner product norm . . . 274
8.4 Examples in $\mathbb{R}^n$ . . . 276
8.5 Examples in the space of square summable sequences . . . 280

8.6 Examples in the space of square integrable functions 282
8.7 Example in the space of random variables with finite second moments . 284
8.8 Orthogonal vectors 285
8.9 Gram-Schmidt procedure 289
8.10 Existence results 296
8.11 Perpendicular spaces 298
8.12 Gram matrix 305
8.13 Exercises 308

9 PROJECTIONS 319
9.1 Closest point counter-examples 320
9.2 Uniqueness of closest point in a subspace 321
9.3 Existence of closest point in a finite dimensional subspace 322
9.4 Orthogonal projection theorem 324
9.5 Projecting onto a 1-dimensional subspace 325
9.6 Approximating e^t 326
9.7 Projecting onto $Col(A)$ 330
9.8 Least squares 333
9.9 Weighted least squares 336
9.10 The dangers of overfitting 342
9.11 Projecting with a general Gram matrix 345
9.12 Fourier series 346
9.13 Minimum mean square error estimation 351
9.14 Projecting onto $Null(A)$ 355
9.15 Projecting onto an affine set 356
9.16 The projection operator 359
9.17 Exercises 360

10 COMPLEX INNER PRODUCT SPACE 379
10.1 Complex conjugate 379
10.2 Hermitian: The conjugate transpose 380
10.3 Complex inner product 381
10.4 Example: The complex inner product space $\mathbb{C}^n$ 383
10.5 Example: Complex inner product space of square integrable functions . 384
10.6 Fourier series via complex exponentials 386
10.7 Are dimensions doubled? 388

IV Eigenvalues, singular values, and their applications 391

11 DETERMINANTS 393

11.1 Axioms of determinants . . . 393
11.2 Existence and uniqueness . . . 395
11.3 Basic consequences . . . 395
11.4 Elementary row operations . . . 398
11.5 Deeper consequences . . . 400
11.6 Determinants by Gaussian elimination . . . 403
11.7 Determinant with block upper triangular structure . . . 403
11.8 Determinant of 2×2 . . . 404
11.9 Determinant of 3×3 and $n \times n$ . . . 405
11.10 Cofactor formulas for $\det(A)$ and A^{-1} . . . 407
11.11 Existence of the determinant function . . . 409
11.12 Geometric properties for $\mathbb{F} = \mathbb{R}$ . . . 410
11.13 Determinant of linear operator . . . 413
11.14 Exercises . . . 414

12 EIGENVALUES 419
12.1 Necessary and sufficient condition . . . 421
12.2 Linear independence property . . . 428
12.3 Eigenvalues of upper triangular matrices . . . 430
12.4 Computing eigenvectors . . . 430
12.5 Spectrum of a matrix . . . 433
12.6 Eigenvalues of a 2×2 matrix . . . 434
12.7 Diagonalizable matrices . . . 436
12.8 Similar matrices . . . 441
12.9 Eigenvalues for linear operators . . . 446
12.10 Exercises . . . 447

13 POLYNOMIALS, ROOTS, AND EIGENVALUES OVER $\mathbb{C}$ 455
13.1 Polynomials over the real field . . . 455
13.2 Polynomials over the complex field . . . 456
13.3 Complex eigenvalues . . . 458
13.4 Eigenvalues for linear operators over field $\mathbb{C}$ . . . 460
13.5 Invariant subspaces . . . 461
13.6 Polynomial functions of matrices . . . 462
13.7 Exercises . . . 466

14 THE ODE $x'(t) = Ax(t)$ 471
14.1 The subspace of solutions . . . 472
14.2 Existence and uniqueness . . . 472

14.3 Solutions from eigenpairs of A . . . 473
14.4 The case when A is diagonalizable . . . 473
14.5 Example: Diagonalizable with real eigenvalues . . . 475
14.6 Example: Diagonalizable with complex eigenvalues . . . 478
14.7 Different solution structure when A is not diagonalizable . . . 482
14.8 Matrix exponential . . . 484
14.9 Exact solution via matrix exponential . . . 488
14.10 Discretizing for nonlinear ODEs . . . 491
14.11 Linear feedback control . . . 494
14.12 Exercises . . . 497

15 Symmetric matrices, Hermitian matrices, Normal matrices 505
15.1 Summary of results . . . 506
15.2 Square matrices have orthonormal columns if and only if orthonormal rows 508
15.3 Diagonalizable form with orthonormal eigenvectors . . . 511
15.4 Spectral Theorem . . . 513
15.5 Properties of real and symmetric matrices . . . 517
15.6 Optimizing $x^\top Bx$ and $|x^\top Bx|$ . . . 519
15.7 Optimizing $||Bx||$ . . . 520
15.8 Example with three distinct eigenvalues . . . 521
15.9 Example with a repeated eigenvalue . . . 523
15.10 Example maximization of $|x^\top Bx|$ . . . 526
15.11 Skew symmetric and skew Hermitian matrices . . . 528
15.12 Additional properties of normal matrices . . . 529
15.13 Exercises . . . 531

16 Singular values 541
16.1 Rank 1 matrices . . . 541
16.2 Singular value decomposition (SVD) . . . 543
16.3 Squared singular values as eigenvalues of $A^\top A$ and $AA^\top$ . . . 547
16.4 SVD for a rank 1 matrix . . . 550
16.5 General procedure for computing SVD . . . 551
16.6 Example . . . 552
16.7 Proof of SVD existence . . . 553
16.8 Maximum expansion factor . . . 555
16.9 Singular values for symmetric matrices . . . 558
16.10 Matrix approximation . . . 559
16.11 Singular values for complex matrices . . . 560
16.12 Singular values for A and A^k . . . 562

16.13 Exercises . . . 563

V Appendices **581**

A Worksheet 1: Speed 583
A.1 Speed 2 . . . 583
A.2 Speed with inner products . . . 583
A.3 Speed with $\mathbb{R}^n$ and $\mathbb{R}^m$ . . . 584
A.4 Solving $Ax = b$ via transposes . . . 584

B Worksheet 2: Pseudoinverse 585
B.1 Null space . . . 585
B.2 One direction . . . 586
B.3 Other direction . . . 586
B.4 Right pseudoinverse . . . 586
B.5 Example . . . 586
B.6 Tall matrices A . . . 587

C Affine sets 589
C.1 Three points . . . 591
C.2 Fields for which $1 + 1 \neq 0$ . . . 593
C.3 Two points . . . 593
C.4 Counter-example when $1 + 1 = 0$ . . . 594
C.5 Intersections . . . 595

D Geometry of $\mathbb{R}^n$ 597
D.1 Row and column perspectives for $Ax = b$ . . . 597
D.2 Hyperplanes . . . 601
D.3 Norm . . . 605
D.4 Rotation matrix for $\mathbb{R}^2$ . . . 609
D.5 Unitary matrices and orthogonality . . . 611
D.6 Realizable rotations in $\mathbb{R}^2$ . . . 613
D.7 Three dimensional rotations . . . 614

E Hamel basis 625
E.1 Definition . . . 625
E.2 Existence of a Hamel basis . . . 628
E.3 Uniqueness property . . . 628
E.4 Existence of inner products for real vector spaces . . . 629
E.5 Infinite dimensional dual space . . . 630

E.6 Things that are not a Hamel basis 631

Bibliography 633

Acknowledgments 635

Preface

These notes are arranged into a book format. The first few chapters treat basic linear algebra concepts with general fields $\mathbb{F}$. This includes matrix operations, linear independence, subspaces, dimension, general vector spaces, and general linear transformations. Finite field applications are considered for binary erasure coding and a graph theory game. Then, the text focuses on the real field $\mathbb{R}$ and considers the theory of projections in general (real) vector spaces. This includes least-squares linear regression, function approximation, and Fourier series. The complex field $\mathbb{C}$ is used for the theory of eigenvalues and singular values. There are a significant number of exercises arranged by chapter.

0.1. Philosophy

Learning linear algebra involves mastering the following four topic areas:

1. Matrices
2. Linear independence
3. Projections
4. Eigenvalues (and singular values)

This text arranges material into these four general areas. Generally speaking: Finite fields work well for $Ax = b$ problems. The real field works well for $Ax \approx b$ problems. The complex field works well for eigenvalue problems.

0.2. Linear independence

The topic of linear independence affects all areas of linear algebra. This text emphasizes the technique of rigorously proving a list of vectors $\{x_1, \ldots, x_n\}$ is linearly independent, using some information given about these vectors, by encouraging students to always

start their proof with the same first line:

"Suppose $\sum_{i=1}^{n} a_i x_i = 0$ for some scalars a_i. We want to show $a_i = 0$ for all i."

Memorizing this first line is, in fact, memorizing the definition of linear independence. It is reasonable to expect both graduate and undergraduate students in a linear algebra class to know the definition of linear independence well enough to give a valid first line of a proof. This text gives a case study on false proofs and gives multiple choice questions that help write first lines for proving $\{x, y, z\}$ are linearly independent, and for proving $\{x_1, ..., x_m, y_1, ..., y_k\}$ are linearly independent (Exercises 3.22 and 3.23). Many students who have no trouble working with three vectors $\{x, y, z\}$ have difficulty with $m + k$ vectors $\{x_1, ..., x_m, y_1, ..., y_k\}$. [No, you are not allowed to leave out the y_j vectors when writing your first line. Yes, you are allowed to call your scalars $a_1, ..., a_m, b_1, ..., b_k$ if you like, or you can call them $a_1, ..., a_{m+k}$ if you prefer.]

This text first develops the linear algebra properties of matrices and vector spaces using a general field $\mathbb{F}$. Once the concepts of span, subspace, linear independence, and dimension have been established in general, this text treats the special case geometry of $\mathbb{R}^n$ (including rotations in 2d and 3d). The fact that two linearly independent vectors in 2-dimensional space must span the space can now be appreciated by illustrating how that works in $\mathbb{R}^2$. The fact that pairwise linear independence does not imply linear independence is illustrated by considering 10 unit-length vectors in $\mathbb{R}^2$, each pointing according to an angle that is two degrees larger than the previous vector: Every pair is linearly independent, but the entire list of 10 cannot be linearly independent since we cannot have 10 linearly independent vectors in 2-dimensional space!

The more someone appreciates that linear independence is a property that applies to a *collection* of vectors, rather than a pairwise property of the vectors, the more that person is surprised by the power of an *inner product*. Remarkably, for the *real field*, linear independence between a collection of vectors can be determined by pairwise inner products via the *Gram matrix*. (It also works for field $\mathbb{C}$.) Also, a surprising theorem shows that any linearly independent vectors $\{x_1, ..., x_n\}$ in a real and finite dimensional vector space can be made to look orthonormal by constructing a suitable inner product (Theorem 8.2 and Corollary 8.2).

0.3. Fields

In modern times, the equation $1 + 1 = 0$ (true in the binary field) is used more by *computers* in just one hour of any given today than the equation $1 + 1 = 2$ (true for the real field) has ever been used by *humans* throughout all of human history. That is because every data packet sent over the internet is encoded and decoded using mod 2 arithmetic via *cyclic*

redundancy check (CRC) codes for error detection.

The first several chapters develop linear algebra for a general field $\mathbb{F}$. The reader can assume $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$ if desired. However, the text considers several important applications of finite field linear algebra to erasure coding for robust communication and distributed data storage (Exercises 5.21, 5.22, 5.23, 5.24, 5.25, 5.26, 5.27, 5.28). The text also shows how the binary field can be used to solve a graph theory game (Exercises 5.17, 5.18, 5.19, 5.20). For pedagogical purposes, finite fields provide an ideal setting to demonstrate the concept of *dimension*. This is because is easy to show that if $\mathbb{F}$ is a finite field with f scalars, any subspace that is spanned by k linearly independent vectors must have exactly f^k vectors. (So *of course* any two bases $\{x_1, ..., x_k\}$ and $\{y_1, ..., y_m\}$ for the same subspace must have $m = k$, since $f^m = f^k$.) The special case examples of finite fields help readers appreciate that similar results hold for infinite fields (such as $\mathbb{R}$ and $\mathbb{C}$), but with necessarily different proofs.

0.4. Gaussian elimination

Gaussian elimination has its own chapter, Chapter 4. This text treats Gaussian elimination as an application of the following result: Suppose we have a matrix A and a vector b and we want to understand the set of solutions to the equation $Ax = b$. If G is an invertible matrix of appropriate size for the multiplication GA to make sense, then

$$\{x : Ax = b\} = \{x : GAx = Gb\}$$

Therefore, left-multiplication of an equation by an invertible matrix G does not change the solution set. (We use "G" to stand for "[G]aussian elimination"). See the related Exercise 4.1 where correct steps lead to a surprising conclusion if we use noninvertible matrices and if we do not properly interpret the meaning of the very first equation.

0.5. Checking answers with computer

For pedagogical purposes, students are encouraged to solve problems by hand as far as possible (this is useful practice for exams). Students are also encouraged to check their answers with numerical solvers. Certain parts of this text give example MATLAB programs, which can easily be translated to other software languages.

0.6. Notation

This text refers to a collection of consecutively numbered equations by stating the first and last equation numbers separated by a hyphen. For example, "equations (8.15)-(8.18)"

refers to the equations (8.15), (8.16), (8.17), (8.18).

If C is a set and x, y, z are objects, we write $x, y, z \in C$ to mean that all three objects x, y, and z are in set C.

Symbols

- $\forall$ means "for all"
- $\implies$ means "implies"
- $\iff$ means "if and only if"

Defining sets

Sets are often defined with braces. For example:

$$C = \{(x_1, x_2, x_3) \in \mathbb{R}^3 : x_1^2 + x_2 = x_3, x_3^2 = 1\}$$

This reads: "C is the set of all (x_1, x_2, x_3) in $\mathbb{R}^3$ such that $x_1^2 + x_2 = x_3$ and $x_3^2 = 1$." In this context, the braces "$\{\cdot\}$" mean "the set of" and the colon symbol ":" means "such that."

Subsets

Suppose C, D are sets and x is an object. We use the following notation:

- $x \in C$ (x is an *element* of C)
- $C \subseteq D$ (C is a *subset* of D, so all elements of C are also in D)
- $C \supseteq D$ (C is a *superset* of D, so $D \subseteq C$)

Two sets C and D are equal, written $C = D$, if and only if $C \subseteq D$ and $D \subseteq C$. The *empty set*, denoted ϕ, is the set with no elements. The empty set is a *subset* of every set. That is, if C is a set then $\phi \subseteq C$. [Some texts use symbol $\subset$ to mean the same as $\subseteq$. We never use the symbol $\subset$.] The empty set is allowed to be an element of a set (such as $A = \{\phi, \textit{red}, 2\}$), but the empty set is not an element of *every* set.

Standard sets

Define the following sets of numbers:

- $\mathbb{N} = \{\text{positive integers}\} = \{1, 2, 3, \ldots\}$
- $\mathbb{Z} = \{\text{integers}\} = \{0, 1, -1, 2, -2, \ldots\}$
- $\mathbb{Q} = \{\text{rational numbers}\} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}$
- $\mathbb{R} = \{\text{real numbers}\}$

The set $\mathbb{N}$ of positive integers is also called the set of *natural numbers*. If C is a nonempty set and $n \in \mathbb{N}$, we define C^n as the set of all n-tuples with entries in C:

$$C^n = \{(c_1, c_2, ..., c_n) : c_i \in C \quad \forall i \in \{1, ..., n\}\}$$

Thus, $\mathbb{R}^n$ denotes the set of n-tuples of real numbers, and $\mathbb{N}^n$ denotes the set of n-tuples of positive integers. For example: $(4, 5, -2) \in \mathbb{R}^3$, $(0, 0.2, 9.9) \in \mathbb{R}^3$, $(1, 100) \in \mathbb{N}^2$.

For $a, b \in \mathbb{R}$ with $a < b$, intervals are defined:

$$[a, b] = \{x \in \mathbb{R} : a \le x \le b\}$$
$$(a, b) = \{x \in \mathbb{R} : a < x < b\}$$
$$[a, b) = \{x \in \mathbb{R} : a \le x < b\}$$
$$(a, b] = \{x \in \mathbb{R} : a < x \le b\}$$

Also

$$[a, \infty) = \{x \in \mathbb{R} : x \ge a\} \quad , \quad (a, \infty) = \{x \in \mathbb{R} : x > a\}$$
$$(-\infty, b] = \{x \in \mathbb{R} : x \le b\} \quad , \quad (-\infty, b) = \{x \in \mathbb{R} : x < b\}$$

For sets of integers, given $a, b \in \mathbb{Z}$ with $a < b$ we use the notation

$$\{a, ..., b\} = \{z \in \mathbb{Z} : a \le z \le b\}$$

0.7. Set or tuple?

Should k vectors be listed as $\{x_1, ..., x_k\}$, or as $(x_1, ..., x_k)$? For example, we might have $k = 3$ and

$$x_1 = (3, 1), x_2 = (9, 9), x_3 = (3, 1)$$

which means some of the vectors in the list can be repeated. Formally, a *k-tuple* is a data structure that allows k items to be listed with possible repetition. Therefore, some texts prefer to write a list of vectors as the tuple $(x_1, ..., x_k)$. Of course, it is common to refer to a single vector in $\mathbb{F}^k$ as $x = (x_1, ..., x_k)$. To emphasize that we are dealing with a list of vectors, rather than a single vector with k components, other texts use notation $\{x_1, ..., x_k\}$. We adopt this latter notation. This means that braces are sometimes used to represent a *tuple* rather than a *set*. (Sets do not care about ordering their elements, and the elements of a set do not have redundancies.) The context makes it clear: If we are dealing with linearly dependent or independent vectors $\{x_1, ..., x_k\}$, we are treating them as a list (the same as a tuple), rather than a set.

PART I.

Scalars and matrices

CHAPTER 1.

Axioms for a field

Let $\mathbb{F}$ be a set of objects that we call *scalars*. Assume $\mathbb{F}$ has at least two distinct scalar objects. Further, suppose there are rules for *addition* and *multiplication* between the scalars so that for any scalars $x \in \mathbb{F}$ and $y \in \mathbb{F}$ we have:

- Scalar addition: $x + y \in \mathbb{F}$
- Scalar multiplication: $xy \in \mathbb{F}$

Multiplication is sometimes represented by $\cdot$ or $\times$ to emphasize the two separate terms, so $xy = x \cdot y = x \times y$. The set of scalars $\mathbb{F}$, together with the rules of addition and multiplication, is said to be a *field* if the following seven properties hold. These properties are called the *axioms for a field*:

1. (Commutative law) For all $x, y \in \mathbb{F}$ we have

$$x + y = y + x$$
$$xy = yx$$

2. (Associative law) For all $x, y, z \in \mathbb{F}$ we have

$$(x + y) + z = x + (y + z)$$
$$(xy)z = x(yz)$$

3. (Distributive law) For all $x, y, z \in \mathbb{F}$ we have

$$x(y+z) = xy + xz$$

4. (Existence of 0) There is a scalar $0 \in \mathbb{F}$, called the *additive identity*, such that $x + 0 = x$ for all $x \in \mathbb{F}$.

5. (Existence of 1) There is a scalar $1 \in \mathbb{F}$ called the *multiplicative identity*, such that $x \cdot 1 = x$ for all $x \in \mathbb{F}$.

6. (Additive inverse) For every $x \in \mathbb{F}$, there is an *additive inverse* $-x \in \mathbb{F}$ such that $x + -x = 0$.

7. (Multiplicative inverse) For every $x \in \mathbb{F}$ such that $x \neq 0$, there is a *multiplicative inverse* $x^{-1} \in \mathbb{F}$ such that $xx^{-1} = 1$.

Basic results for fields (such as $x \cdot 0 = 0$ and $-(-x) = x$ for all $x \in \mathbb{F}$) are proven in Section 1.5. Scalars can be viewed as generalizations of *numbers*. Common scalar fields are:

- The real numbers $\mathbb{R}$ with standard arithmetic.
- The rational numbers $\mathbb{Q}$ with standard arithmetic.
- The complex numbers $\mathbb{C}$ with complex arithmetic.
- The finite field $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic ($1 + 1 = 0$).
- The finite field $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ for a prime number p and with mod p arithmetic ($1 + (p-1) = 0$).

These fields are discussed in more detail in Section 1.4. The set of integers $\mathbb{Z}$ with standard arithmetic satisfies Axioms 1-6, but does *not* form a field because it fails the multiplicative inverse axiom (Axiom 7). Specifically, $2 \in \mathbb{Z}$, but 2 has no multiplicative inverse in $\mathbb{Z}$. That is, there is no integer $z \in \mathbb{Z}$ such that $2z = 1$.

1.1. Summing in any order

Let $\mathbb{F}$ be a field. Let x, y, z be elements of $\mathbb{F}$. By repeated application of commutativity and associativity we have

$$(x+y)+z = (y+x)+z = y+(x+z) = y+(z+x) = (y+z)+x$$

and so on. Thus, summing the three elements x, y, z in any order produces the same result. Therefore, we can write this sum without parentheses

$$x + y + z$$

since any way of placing parentheses gives the same answer.

By induction, it can formally be proven that for all $n \in \mathbb{N}$, if we have a collection of scalars $x_1, \ldots, x_n$ in some field $\mathbb{F}$, then summing these scalars in any order produces the same result. We use the convenient summation notation

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \cdots + x_n$$

We do not need to use parentheses to specify how these n scalars are summed. For example, with $n = 4$ we have

$$\sum_{i=1}^{4} x_i = x_1 + (x_2 + (x_3 + x_4)) = (x_1 + x_2) + (x_3 + x_4) = x_1 + x_2 + (x_3 + x_4)$$

Also

$$\sum_{i=1}^{4} x_i = x_1 + x_2 + x_3 + x_4 = x_2 + x_3 + x_1 + x_4 = x_3 + x_1 + x_4 + x_2$$

and so on.

A particularly interesting application of the fact that a finite number of scalars can be summed in any order is the *sum switching rule* for doubly indexed scalars: Fix some field $\mathbb{F}$. Fix $m, n \in \mathbb{N}$. Suppose we have a collection of mn scalars $a_{ij} \in \mathbb{F}$ for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$. We can organize them into a *matrix* with m rows and n columns. For example, with $m = 3$, $n = 4$ and $\mathbb{F} = \mathbb{R}$, our matrix might be this:

$$\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 2 & 10 \\ 0 & 4 & 1 & 1 \\ 2 & 2 & 2 & 5 \end{bmatrix} \quad \text{[sum of all entries is 30]} \tag{1.1}$$

One way to sum over all a_{ij} in the matrix is to first find the *row sum* for each row $i \in \{1, \ldots, m\}$ (call each such value $RowSum_i$), then sum $RowSum_i$ over all i:

$$sum = \sum_{i=1}^{m} \underbrace{\left(\sum_{j=1}^{n} a_{ij} \right)}_{RowSum_i}$$

Another way is to first find the *column sum* for each column $j \in \{1, \ldots, n\}$ (call each such value $ColSum_j$), then sum $ColSum_j$ over all j:

$$sum = \sum_{j=1}^{n} \underbrace{\left(\sum_{i=1}^{m} a_{ij} \right)}_{ColSum_j}$$

Since both ways lead to the same answer, we have the famous *sum switching rule*, which holds for summing doubly indexed scalars a_{ij} in any field $\mathbb{F}$:

$$\sum_{i=1}^{m} \left(\sum_{j=1}^{n} a_{ij} \right) = \sum_{j=1}^{n} \left(\sum_{i=1}^{m} a_{ij} \right)$$

The parentheses in the above formula can be distracting. If we agree to compute the inside sums first, we can write the sum switching rule by removing the parentheses:

(1.2)
$$\boxed{\sum_{i=1}^{m} \sum_{j=1}^{n} a_{ij} = \sum_{j=1}^{n} \sum_{i=1}^{m} a_{ij}} \quad \text{[sum switching rule]}$$

For example, using field $\mathbb{R}$ and the matrix in (1.1), and using R_i and C_j as shorthand for $RowSum_i$ and $ColSum_j$, we have

$$\left[\begin{array}{cccc|c} 1 & 0 & 2 & 10 & R_1 = 13 \\ 0 & 4 & 1 & 1 & R_2 = 6 \\ 2 & 2 & 2 & 5 & R_3 = 11 \\ \hline C_1 = 3 & C_2 = 6 & C_3 = 5 & C_4 = 16 & sum = 30 \end{array}\right]$$

Then

$$\sum_{i=1}^{3} R_i = 13 + 6 + 11 = 30$$
$$\sum_{j=1}^{4} C_j = 3 + 6 + 5 + 16 = 30$$

1.2. Multiplying in any order

Let $\mathbb{F}$ be a field. Commutativity and associativity also hold for multiplication. So we can multiply a finite collection of scalars in any order. If $x, y, z \in \mathbb{F}$ then

$$x(yz) = x(zy) = (xz)y = y(xz)$$

and so on. We can use notation xyz without the parentheses. Fix $n \in \mathbb{N}$. If $x_1, \ldots, x_n$ are a collection of scalars in $\mathbb{F}$ then we define the product of all of them as

$$\prod_{i=1}^{n} x_i = x_1 x_2 \cdots x_n$$

and we do not need to specify parentheses.

Fix $m, n \in \mathbb{N}$. If a_{ij} are doubly indexed scalars in $\mathbb{F}$ for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$ then the *product switching rule* is

$$\prod_{i=1}^{m} \prod_{j=1}^{n} a_{ij} = \prod_{j=1}^{n} \prod_{i=1}^{m} a_{ij}$$

where we avoid parentheses in this formula by agreeing to take the inside products first.

1.3. Uniqueness of additive and multiplicative inverses

Axiom 6 only requires each scalar in a field to have *at least one* additive inverse. Can there be more than one? No: Suppose $\mathbb{F}$ is a field. Fix $x \in \mathbb{F}$. Suppose there are two distinct scalars y and z in $\mathbb{F}$ that *both* satisfy the requirements of an additive inverse for x:

$$x + y = 0 \tag{1.3}$$

$$x + z = 0 \tag{1.4}$$

Since y and z are assumed to be distinct, we know $y \neq z$. We now reach a contradiction:

$$\begin{aligned}
y &= y + 0 && \text{[Axiom 4: 0 is the additive identity]} \\
&= y + (x + z) && \text{[Substitute (1.4)]} \\
&= (y + x) + z && \text{[Axiom 2: Moving the parentheses]} \\
&= (x + y) + z && \text{[Axiom 1: Commutativity]} \\
&= 0 + z && \text{[Substitute (1.3)]} \\
&= z + 0 && \text{[Axiom 1: Commutativity]} \\
&= z && \text{[Axiom 4: 0 is the additive identity]}
\end{aligned}$$

and so $y = z$, contradicting the fact $y \neq z$. Thus, there is one and only one additive inverse for x. This unique additive inverse for x is denoted $-x$. Exercise 1.4 shows that multiplicative inverses are also unique.

1.4. Example fields

1.4.1. Real numbers

The set of real numbers $\mathbb{R}$, together with standard addition and multiplication, is assumed to form a field. [With this assumption, the set $\mathbb{Q}$ of rational numbers can be verified to be a field by checking *closure* for rational sum and multiply operations, see Exercise 1.22.]

1.4.2. Complex numbers

Define a special object i such that $i \notin \mathbb{R}$. Define $\mathbb{C}$, called the set of *complex numbers*, as the set of all objects of the form $a + bi$ where $a, b \in \mathbb{R}$. Currently, this notation just acts to separate a and b, so $\mathbb{C}$ can be equivalently defined as the set of all ordered pairs (a, b) such that $a, b \in \mathbb{R}$. The rules of scalar multiplication are defined as follows: For scalars $z, w \in \mathbb{C}$, where $z = a + bi$ and $w = c + di$ for some $a, b, c, d \in \mathbb{R}$, define

- Addition: $(a + bi) + (c + di) = (a + c) + (b + d)i$
- Multiplication: $(a + bi)(c + di) = (ac - bd) + (ad + bc)i$

We can treat every real number as a special case of a complex number with the definition

$$x \in \mathbb{R} \implies x = x + 0i \in \mathbb{C}$$

We can also treat the special symbol i itself as a scalar in $\mathbb{C}$ with the definition

$$i = 0 + 1i \in \mathbb{C}$$

Then the rule of multiplication implies $i^2 = -1$ because

$$\begin{aligned} i^2 &= (0 + 1i)(0 + 1i) \\ &\overset{(a)}{=} (0 \cdot 0 - 1 \cdot 1) + (0 \cdot 1 + 1 \cdot 0)i \\ &= -1 + 0i \\ &= -1 \end{aligned}$$

where equality (a) uses the definition of scalar multiplication for $\mathbb{C}$. This property of having a scalar that satisfies $i^2 = -1$ is the motivation behind the definition of scalar multiplication

given above. We now verify the 7 axioms hold for $\mathbb{C}$. Our proof assumes we already know the 7 axioms hold for $\mathbb{R}$.

- Axioms 1-3: To show commutativity holds (Axiom 1), fix $z, w \in \mathbb{C}$. We first show $z + w = w + z$. We know $z = a + bi$ and $w = c + di$ for some real numbers a, b, c, d. By the rules of scalar addition for $\mathbb{C}$ we have

$$\begin{aligned} z + w &= (a + bi) + (c + di) \\ &= (a + c) + (b + d)i \\ &\overset{(a)}{=} (c + a) + (d + b)i \\ &= (c + di) + (a + bi) \\ &= w + z \end{aligned}$$

  where the key step (a) uses the assumption that addition for real numbers is commutative. We now show $zw = wz$. The rules of multiplication in $\mathbb{C}$ yield

$$\begin{aligned} zw &= (a + bi)(c + di) \\ &= (ac - bd) + (ad + bc)i \\ &\overset{(a)}{=} (ca - db) + (cb + da)i \\ &= (c + di)(a + bi) \\ &= wz \end{aligned}$$

  where the key step (a) uses the assumption that addition and multiplication of real numbers is commutative (for example, it replaces ac with ca, and replaces bd with db). This proves Axiom 1 holds. The proofs of Axioms 2 and 3 are similar (see Exercise 1.21).

- Axioms 4-5: To show Axioms 4 and 5 hold, we define our candidates for additive identity and multiplicative identity by

$$\begin{aligned} 0 &= 0 + 0i \\ 1 &= 1 + 0i \end{aligned}$$

  Now fix $z \in \mathbb{C}$. Then $z = a + bi$ for some $a, b \in \mathbb{R}$. We have by our rules for addition and multiplication in $\mathbb{C}$:

$$\begin{aligned} z + 0 &= (a + bi) + (0 + 0i) = (a + 0) + (b + 0)i = a + bi = z \\ z \cdot 1 &= (a + bi) \cdot (1 + 0i) = (a \cdot 1 - b \cdot 0) + (a \cdot 0 + b \cdot 1)i = a + bi = z \end{aligned}$$

so Axioms 4 and 5 hold, with 0 and 1 indeed working as desired.

- Axiom 6: Fix $z \in \mathbb{C}$. We want to show there is a scalar, call it $-z$, that satisfies $z + -z = 0$. Since $z = a + bi$ for some $a, b \in \mathbb{R}$, let us try the following candidate for $-z$:

$$-z = (-a) + (-b)i$$

This works:

$$\begin{aligned}(a + bi) + (-a + (-b)i) &= (a + -a) + (b + -b)i \\ &= 0 + 0i \\ &= 0\end{aligned}$$

Therefore, Axiom 6 holds. We do not need to prove $-a + (-b)i$ is the *unique* additive inverse of $a + bi$. Once the other axioms are proven, uniqueness is guaranteed by the result of Section 1.3.

- Axiom 7: Fix $z \in \mathbb{C}$ such that $z \neq 0$. We need to show there is some scalar, call it z^{-1}, that satisfies $zz^{-1} = 1$. We know $z = a + bi$ for some $a, b \in \mathbb{R}$. Since $z \neq 0$, we know that

$$(a, b) \neq (0, 0)$$

That is, either $a \neq 0$, or $b \neq 0$ (or both). Let $z^{-1} = c + di$. We want to design real numbers c, d such that

$$(a + bi)(c + di) = 1$$

It turns out that such c, d can be constructed. The construction relies on the assumption that either $a \neq 0$ or $b \neq 0$. This design of c, d is left to Exercise 1.7.

1.4.3. The binary field $\mathbb{F}_2 = \{0, 1\}$

Define $\mathbb{F}_2 = \{0, 1\}$. In this field there are only two scalars, namely, 0 and 1. Addition and multiplication are defined by mod 2 arithmetic:

$$\begin{aligned} 0 + 0 &= 0 & \qquad 0 \cdot 0 &= 0 \\ 0 + 1 &= 1 & \qquad 0 \cdot 1 &= 0 \\ 1 + 0 &= 1 & \qquad 1 \cdot 0 &= 0 \\ 1 + 1 &= 0 & \qquad 1 \cdot 1 &= 1 \end{aligned}$$

These rules can be equivalently written as addition and multiplication tables:

+	0	1
0	0	1
1	1	0

×	0	1
0	0	0
1	0	1

The equation $1 + 1 = 0$ (which is true for mod 2 arithmetic) is used much more by *computers* in just one hour of a day than the equation $1 + 1 = 2$ (which is true for real numbers) has ever been used by *humans* throughout all of human history. This is because mod 2 arithmetic is built into the cyclic redundancy check (CRC) codes that are used to detect errors in every packet of data that is sent over the internet. It can be verified that the axioms for a field hold for $\mathbb{F}_2$:

- Axiom 1 (commutativity): We can use the given arithmetic tables to verify that $x + y = y + x$ and $xy = yx$ is true for all four possible values of x and y:

  $$(x, y) \in \{(0,0), (0,1), (1,0), (1,1)\}$$

  Alternatively, we can observe that commutativity must hold because the addition and multiplication tables are *symmetric*.

- Axioms 2-3: We can test all eight combinations of $(x, y, z) \in \{0,1\}^3$ under the given arithmetic tables to verify the following always hold:

  $$\begin{aligned} (x + y) + z &= x + (y + z) \\ (xy)z &= x(yz) \\ x(y + z) &= xy + xz \end{aligned}$$

  We can write a computer program to test these (see MATLAB program given below). Here is a proof of the distributive law that does not need a computer: Fix $x, y, z \in \{0,1\}$. If $x = 0$ then

  $$x(y+z) = 0(y+z) \overset{(a)}{=} 0 \overset{(b)}{=} 0 + 0 \overset{(c)}{=} 0 \cdot y + 0 \cdot z \overset{(d)}{=} x \cdot y + x \cdot z$$

  where (a) holds because the multiplication table shows that 0 times anything gives zero; (b) holds because the addition table shows $0 + 0 = 0$; (c) holds again because 0 times anything is zero; (d) holds by substituting $x = 0$. On the other hand, if $x = 1$ then

  $$x(y+z) = 1(y+z) \overset{(a)}{=} y + z \overset{(b)}{=} 1 \cdot y + 1 \cdot z \overset{(c)}{=} xy + xz$$

  where (a) holds because the multiplication table shows that 1 times anything gives that same thing back; (b) holds for the same reason; (c) substitutes $x = 1$.

- Axioms 4-5: To show $x + 0 = x$ and $x \cdot 1 = x$, we just test the two cases $x \in \{0, 1\}$ using the arithmetic tables:

$$0 + 0 = 0$$
$$1 + 0 = 1$$
$$0 \cdot 1 = 0$$
$$1 \cdot 1 = 1$$

- Axiom 6: It is easy to use the addition table to verify $-x = x$ for all $x \in \{0, 1\}$:

$$0 + 0 = 0 \quad [\text{so } -0 = 0]$$
$$1 + 1 = 0 \quad [\text{so } -1 = 1]$$

- Axiom 7: From the multiplication table we have $1 \cdot 1 = 1$, so $1^{-1} = 1$.

MATLAB program to verify Axioms 2-3

Listing 1.1. MATLAB for testing Axioms 2-3 with mod 2 field

```
%This program tests Axioms 2-3 for mod-2 arithmetic

global AddTable
global MultTable

AddTable= [0 1;1 0];
MultTable = [0 0;0 1];

error = 0;

%%% LOOP
for x=0:1
for y=0:1
for z=0:1

%Test Axiom 2
if ADD(ADD(x,y),z)~= ADD(x,ADD(y,z))
    error = 1;
end

if MULT(MULT(x,y),z) ~= MULT(x,MULT(y,z))
    error = 1;
end
```

```
%Test Axiom 3
if MULT(x,ADD(y,z)) ~= ADD(MULT(x,y),MULT(x,z))
    error = 1;
end

end
end
end
%%%% END LOOP

error

%The ADD and MULT functions
function w=ADD(x,y)
  global AddTable
  w=AddTable(x+1,y+1);
end

function w=MULT(x,y)
  global MultTable
  w = MultTable(x+1,y+1);
end
```

1.4.4. The trinary field $\mathbb{F}_3 = \{0, 1, 2\}$

Define $\mathbb{F}_3 = \{0, 1, 2\}$. In this field there are only three scalars: The two scalars 0 and 1 that are needed for *every* field, and an additional scalar called 2. Addition and multiplication are defined by mod 3 arithmetic, given by the following tables:

+	0	1	2
0	0	1	2
1	1	2	0
2	2	0	1

×	0	1	2
0	0	0	0
1	0	1	2
2	0	2	1

Axiom 1 (commutativity) holds because the addition and multiplication tables are symmetric. Axioms 2-3 can be verified by modifying the computer program given in the previous subsection (see Exercise 1.19). Axioms 4-7 can be verified in a manner similar to the case for mod 2 arithmetic in the previous subsection.

One important property that holds for all fields is this: *Every scalar appears exactly once in each row and column of the addition table.* Why? This question is explored in Exercise 1.14.

The same is *almost true* for the multiplication table (see Exercise 1.15). The only exception is the first row and column: These fail because 0 times anything is 0. The fact $0 \cdot x = 0$ is *not* an axiom, but is proven from the axioms in Section 1.5.

1.4.5. The finite field $\mathbb{F}_p$ for p prime

Fix p as a prime number:

$$p \in \{2, 3, 5, 7, 11, \ldots\}$$

Define $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$. For $x, y \in \mathbb{F}_p$ define scalar addition and multiplication according to mod p arithmetic of integers. [Readers who are unfamiliar with modular arithmetic can read a quick primer on that subject in Section 1.7.] For all $x, y \in \mathbb{F}_p$ we have

- Addition: $x + y = \mod((x + y)_{\mathbb{Z}}, p)$
- Multiplication: $xy = \mod((xy)_{\mathbb{Z}}, p)$

where $(x + y)_{\mathbb{Z}}$ and $(xy)_{\mathbb{Z}}$ denote the standard addition and multiplication of integers. Specifically, recall that every integer n can be written $n = kp + r$ for some unique integers k, r such that $r \in \{0, 1, \ldots, p-1\}$, and for such n we define

$$\mod(n, p) = r$$

That is, for each integer n, we define $\mod(n, p)$ as the *remainder* $r \in \{0, \ldots, p-1\}$ when we divide n by p.

For any integer $p \geq 2$ (possibly not prime), it can be shown that mod p arithmetic satisfies Axioms 1-6. However, the multiplicative inverse property (Axiom 7) holds for mod p arithmetic if and only if p is a prime number (this is formally proven in Section 1.7). Therefore, the set $\mathbb{F}_p$ with mod p arithmetic is a field if and only if p is prime. The cases $p = 2$ and $p = 3$ have been discussed in previous subsections. The next field with mod p arithmetic occurs when $p = 5$, it is $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ with mod 5 arithmetic.

There exists a finite field of size 4, but its scalars $\{0, 1, a, b\}$ use arithmetic tables that are different from mod 4 arithmetic (see Exercise 1.17). A famous result by mathematicians Galois and Moore is that a finite field of size q exists if and only if $q = p^n$ for some prime p and some $n \in \mathbb{N}$ (such fields only use mod q arithmetic when $n = 1$). Fields with size 2^n can be used to efficiently build *erasure codes* for binary data (discussed in a later chapter).

1.5. Basic results for fields

Let $\mathbb{F}$ be any set which, together with rules for addition and multiplication, satisfies the axioms for a field.

LEMMA 1.1. $-0 = 0$.

PROOF. The scalar -0 exists by Axiom 6. By definition of -0 we have

$$\begin{aligned} 0 &= 0 + -0 && \text{[definition of } -0 \text{ (Axiom 6)]} \\ &= -0 + 0 && \text{[commutativity (Axiom 1)]} \\ &= -0 && \text{[0 is the additive identity (Axiom 4)]} \end{aligned}$$

□

LEMMA 1.2. $x \cdot 0 = 0$ *for all* $x \in \mathbb{F}$.

PROOF. Fix $x \in \mathbb{F}$. We have

$$\begin{aligned} x &\overset{(a)}{=} x \cdot 1 \\ &\overset{(b)}{=} x \cdot (1 + 0) \\ &\overset{(c)}{=} x \cdot 1 + x \cdot 0 \\ &\overset{(d)}{=} x + x \cdot 0 \end{aligned}$$

where (a) uses 1 as the multiplicative identity; (b) uses $1 + 0 = 1$; (c) uses the distributive law; (d) again uses 1 as the multiplicative identity. Therefore,

$$x = x + x \cdot 0 \tag{1.5}$$

Adding $-x$ to both sides of (1.5) gives

$$\begin{aligned} -x + x &= -x + (x + x \cdot 0) \\ &= (-x + x) + x \cdot 0 \quad \text{[by moving parentheses (Axiom 2)]} \end{aligned}$$

Thus

$$\underbrace{-x + x}_{0} = \underbrace{-x + x}_{0} + x \cdot 0$$

The definition of $-x$ ensures $-x + x = 0$, so the above equality simplifies to $0 = x \cdot 0$. □

The above proof used a step that adds the same thing to both sides of an equation. Formally, this is justified by the *substitution rule*: If a, b, c are scalars and if $a = b$, then we can replace b with a in any expression without changing the value of that expression. So

the expression $c + a$ is the same as $c + b$. That is

$$(a = b) \implies c + a = c + b$$

which is the same as adding c to both sides of the equation $a = b$. Likewise, the expression ca is the same as cb and so

$$(a = b) \implies ca = cb$$

which is the same as multiplying both sides of $a = b$ by c. The substitution rule is not part of the axioms for a field because it is a fundamental property about the nature of equality between two objects. The substitution rule holds more generally when performing the same operation to both sides of an equation.

LEMMA 1.3. $1 \neq 0$.

PROOF. Suppose $1 = 0$ (we reach a contradiction). Since we have assumed $\mathbb{F}$ has at least two distinct elements, there must be some element $x \in \mathbb{F}$ such that $x \neq 0$. Then $x = x \cdot 1 = x \cdot 0 = 0$, contradiction. □

LEMMA 1.4. $1^{-1} = 1$.

PROOF. See Exercise 1.2. □

LEMMA 1.5. $-(-x) = x$ *and* $-x = (-1)x$ *for all* $x \in \mathbb{F}$.

PROOF. Fix $x \in \mathbb{F}$. Then $-x \in \mathbb{F}$ and so

$$-x + -(-x) = 0$$

Adding x to both sides gives

$$(x + -x) + -(-x) = x + 0$$

Simplifying gives $-(-x) = x$. The proof that $-x = (-1)x$ is Exercise 1.3. □

As a special case of the above lemma, we obtain $(-1)(-1) = 1$ because

$$1 = -(-1) = (-1)(-1)$$

This, together with Lemma 1.5, implies $(-x)(-y) = xy$ for all $x, y \in \mathbb{F}$.

LEMMA 1.6. *If $x \in \mathbb{F}$ and $x \neq 0$, then $x^{-1} \neq 0$ and $(x^{-1})^{-1} = x$.*

PROOF. Fix $x \in \mathbb{F}$, $x \neq 0$. The multiplicative inverse axiom ensures x^{-1} exists as a scalar in $\mathbb{F}$. If $x^{-1} = 0$ then we have

$$1 = xx^{-1} = x0 = 0$$

a contradiction. So we must have $x^{-1} \neq 0$. Then x^{-1} itself has a multiplicative inverse and

$$x^{-1}(x^{-1})^{-1} = 1$$

Multiplying both sides by x gives the result. □

LEMMA 1.7. *If $x, y \in \mathbb{F}$ then*

$$xy = 0 \iff (x = 0 \text{ or } y = 0)$$

PROOF. See Exercise 1.1. □

1.6. Subtraction and division

The operations of scalar addition and multiplication are fundamental for a field. Subtraction and division are *defined* by addition and multiplication together with the additive inverse and multiplicative inverse axioms (see also Exercise 1.12):

1. For $x, y \in \mathbb{F}$ define
$$x - y = x + (-y)$$
2. For $x, y \in \mathbb{F}$ with $y \neq 0$ define
$$\frac{x}{y} = xy^{-1}$$

Given $x, y \in \mathbb{F}$, the famous add-subtract technique is

$$x = x + y - y$$

1.7. Modular arithmetic

DEFINITION 1.1. *Fix an integer $d \geq 2$. For each integer n, we define $\text{mod}(n, d)$ as the unique integer $r \in \{0, 1, \ldots, d-1\}$ such that $n = kd + r$ for some integer k. This value r is the* remainder *when the integer n is divided by the integer d.*

It follows that $\text{mod}\,(n, d) = 0$ if and only if n is an integer multiple of d. To compute $\text{mod}\,(n, d)$, we start with n and then add or subtract multiples of d until the result is in

the set $\{0, \ldots, d-1\}$. For example, with $d = 5$ we have

$$\begin{aligned}
\text{mod}\,(3,5) &= 3 \quad [\text{since } 3 = 0(5) + 3] \\
\text{mod}\,(21,5) &= 1 \quad [\text{since } 21 = 4(5) + 1] \\
\text{mod}\,(22,5) &= 2 \quad [\text{since } 22 = 4(5) + 2] \\
\text{mod}\,(100,5) &= 0 \quad [\text{since } 100 = 20(5) + 0] \\
\text{mod}\,(-3,5) &= 2 \quad [\text{since } -3 = -1(5) + 2]
\end{aligned}$$

With $d = 8$ we have

$$\begin{aligned}
\text{mod}\,(3,8) &= 3 \quad [\text{since } 3 = 0(8) + 3] \\
\text{mod}\,(21,8) &= 5 \quad [\text{since } 21 = 2(8) + 5] \\
\text{mod}\,(22,8) &= 6 \quad [\text{since } 22 = 2(8) + 6] \\
\text{mod}\,(100,8) &= 4 \quad [\text{since } 100 = 12(8) + 4] \\
\text{mod}\,(-3,8) &= 5 \quad [\text{since } -3 = -1(8) + 5]
\end{aligned}$$

These can be verified by the MATLAB mod $(\cdot)$ command:

Listing 1.2. MATLAB mod command

```
>> mod(3,5)

ans =

     3

>> mod(21,5)

ans =

     1
```

1.7.1. Equivalent mod d

Fix $d \geq 2$ as an integer. Two integers a and b are said to be *equivalent (mod d)*, written

$$a = b \pmod{d}$$

if a and b differ by an integer multiple of d, that is, $a - b = kd$ for some integer k. It follows that a and b are equivalent (mod d) if and only if

$$\mathrm{mod}(a, d) = \mathrm{mod}(b, d)$$

An equation of the type

$$ab + bc^2 = ad + b^2a^3 + a \pmod{d}$$

with some integers a, b, c, means the left-hand-side and right-hand-side differ by an integer multiple of d, so $(ab + bc^2) - (ad + b^2a^3 + a) = kd$ for some integer k. The next lemma shows that if we have an expression involving sums and products of integers, then any integer in the expression can be replaced with an equivalent integer (where equivalence is in the mod d sense) without changing the mod d value of that expression.

LEMMA 1.8. *Fix $d \geq 2$ as an integer. Let a, b, c, k be integers. Then*

(1.6) $$a + b = a + (b + kd) \quad (mod\ d)$$

(1.7) $$a + bc = a + b(c + kd) \quad (mod\ d)$$

(1.8) $$a + bc = mod(a, d) + mod(b, d)mod(c, d) \quad (mod\ d)$$

PROOF. Equality (1.6) is true because $a + b$ and $a + (b + kd)$ differ by an integer multiple of d:

$$(a + (b + kd)) - (a + b) = kd = \text{integer multiple of } d$$

The proof of equality (1.7) is similar. Equality (1.8) holds because

$$\mathrm{mod}(a, d) = a - k_1 d$$
$$\mathrm{mod}(b, d) = b - k_2 d$$
$$\mathrm{mod}(c, d) = c - k_3 d$$

for some particular integers k_1, k_2, k_3. So proving (1.8) reduces to proving

$$a + bc = (a - k_1 d) + (b - k_2 d)(c - k_3 d) \pmod{d}$$

This is true: $a + bc$ and $(a - k_1 d) + (b - k_2 d)(c - k_3 d)$ differ by an integer multiple of d. □

By using the above lemma with induction, we can compute the sum or product of any finite collection of integers in the mod d sense by first finding the mod d values of those

integers. Indeed, fix $d \geq 2$ as an integer. Let $n_1, ..., n_k$ be a collection of integers. Define

$$\overline{n}_i = \text{mod}(n_i, d) \quad \forall i \in \{1, ..., k\}$$

Then

$$n_1 + n_2 + \cdots + n_k = \overline{n}_1 + \overline{n}_2 + \cdots + \overline{n}_k \pmod{d}$$
$$n_1 n_2 \cdots n_k = \overline{n}_1 \overline{n}_2 \cdots \overline{n}_k \pmod{d}$$

Also, we have, for example,

$$n_1 + n_1 n_2^5 n_3 + n_3 n_4 n_5^2 = \overline{n}_1 + \overline{n}_1 \overline{n}_2^5 \overline{n}_3 + \overline{n}_3 \overline{n}_4 \overline{n}_5^2 \pmod{d}$$

Given an integer $d \geq 2$, these rules can show the set $\{0, 1, ..., d-1\}$, together with mod d arithmetic, satisfies Axioms 1-6 of a field. Specifically, modular arithmetic is commutative, associative, and distributive. Also, 0 and 1 serve as the additive and multiplicative identities. Further, for every $x \in \{0, 1, ..., d-1\}$ we can define $-x = \text{mod}((-x)_{\mathbb{Z}}, d)$, where $(-x)_{\mathbb{Z}}$ refers to the negative integer and $-x$ refers to the scalar in $\{0, 1, ..., d-1\}$ that satisfies $x + -x = 0$ in the mod d sense. However, the multiplicative inverse axiom (Axiom 7) holds if and only if d is a prime number. Therefore, $\mathbb{F}_d = \{0, 1, ..., d-1\}$ with mod d arithmetic is a field if and only if d is prime. This is explained in the next two subsections. The next lemma is a useful consequence of modular arithmetic.

LEMMA 1.9. *Let n be a positive integer. Express n in its base 10 expansion:*

$$n = d_k d_{k-1} \cdots d_1 d_0 = \sum_{i=0}^{k} d_i 10^i$$

for some integer $k \geq 0$. Then

a) n is divisible by 9 if and only if $\sum_{i=0}^{k} d_i$ is divisible by 9. For example, 78303015 is divisible by 9 because $7 + 8 + 3 + 3 + 1 + 5 = 27$.

b) n is divisible by 3 if and only if $\sum_{i=0}^{k} d_i$ is divisible by 3. For example, 1893000372 is divisible by 3 but not by 9 because $1 + 8 + 9 + 3 + 3 + 7 + 2 = 33$.

PROOF. We prove only part (a) (the proof of (b) is similar). Note that $\text{mod}(10, 9) = 1$. Thus, in any expression that takes sums and products of integers, every instance of 10 can be replaced by 1 without affecting the mod 9 value of the expression. So

$$n = \sum_{i=0}^{k} d_i (10)^i = \sum_{i=0}^{k} d_i (1)^i = \sum_{i=0}^{k} d_i \pmod{9}$$

Thus, $n = 0 \pmod 9$ if and only if $\sum_{i=0}^{k} d_i = 0 \pmod 9$. □

1.7.2. Axiom 7 fails when p is not prime

It is relatively easy to see that Axiom 7 fails for mod p arithmetic when p is *not* a prime number. For example, suppose $p = 6$ and consider the set of distinct scalars $\{0, 1, 2, 3, 4, 5\}$ with mod 6 arithmetic. Suppose this forms a field (we reach a contradiction). Since $2 \cdot 3 = 6$ we have

$$2 \cdot 3 = 0 \pmod 6 \tag{1.9}$$

However, we know that $2 \neq 0$ and $3 \neq 0$ (recall that $\{0, 1, 2, 3, 4, 5\}$ are distinct scalars). Since $2 \neq 0$, and since we assume we have a field, there must be some scalar $2^{-1} \in \{0, 1, 2, 3, 4, 5\}$ for which

$$2^{-1} \cdot 2 = 1 \pmod 6$$

Multiplying both sides of (1.9) by 2^{-1} gives

$$\underbrace{(2^{-1} \cdot 2)}_{1} \cdot 3 = 0 \pmod 6$$

which means $1 \cdot 3 = 0 \pmod 6$, that is, $3 = 0$, a contradiction.

More generally, let $p \geq 2$ be an integer that is not prime. Suppose Axiom 7 holds for $\{0, 1, \ldots, p-1\}$ with mod p arithmetic (we reach a contradiction). Then $\{0, 1, \ldots, p-1\}$ with mod p arithmetic forms a field. Since p is not prime, there must be two nonzero integers $a, b \in \{1, 2, \ldots, p-1\}$ such that $ab = p$ (in the usual sense of multiplication of real numbers), and so

$$ab = 0 \pmod p$$

Then, a similar argument reaches a contradiction.

1.7.3. Axiom 7 holds when p is prime

LEMMA 1.10. *Fix p as a prime number.*

a) If $a, b \in \mathbb{Z}$ then

$$(ab = 0 \pmod p) \iff (mod(a, p) = 0 \text{ or } mod(b, p) = 0)$$

b) If $a, a_1, \ldots, a_k \in \mathbb{Z}$ (where k is some positive integer) then

$$\left(\prod_{i=1}^{k} a_i = 0 \pmod p\right) \iff (mod(a_i, p) = 0 \quad \text{for some } i \in \{1, \ldots, k\})$$

$$(a^k = 0 \pmod p) \iff (a = 0 \pmod p)$$

PROOF. For the forward direction of part (a), suppose $ab = 0 \pmod{p}$. Then

$$ab = kp$$

for some integer k. Suppose neither a nor b are equivalent to 0 mod p (we reach a contradiction). It means that neither a nor b is a multiple of p. Consider taking a prime factorization of both sides of the above equation:

$$\underbrace{a}_{\text{(no } p \text{ factors)}} \quad \underbrace{b}_{\text{(no } p \text{ factors)}} = \underbrace{kp}_{\text{contains a factor } p}$$

It is impossible for the right-hand-side to have a prime factor p when the left-hand-side does not. This is the contradiction.

For the reverse direction of part (a), suppose $\mod(a, p) = 0$. We want to show ab is equivalent to 0 (mod p). Since $\mod(a, p) = 0$, we know $a = kp$ for some integer k and so

$$ab = kpb$$

so ab is an integer multiple of p, meaning that ab is equivalent to 0 (mod p) and we are done. The case $\mod(b, p) = 0$ is similar. This completes the proof of part (a).

The proof of part (b) follows from part (a) by induction. □

Now let p be prime. Define $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ and consider mod p arithmetic. We want to show this forms a field. We already know Axioms 1-6 hold (they hold even if p is not prime, provided that $p \geq 2$). It suffices to show Axiom 7 (every nonzero scalar has a multiplicative inverse). Fix $z \in \mathbb{F}_p$ such that $z \neq 0$. Then $z \in \{1, \ldots, p-1\}$. We want to show there is a scalar $s \in \mathbb{F}_p$ such that $zs = 1$. Consider the sequence of powers of z

$$z, z^2, z^3, z^4, \ldots \tag{1.10}$$

Each of the z^j numbers in this sequence has a mod p value in the set $\{0, \ldots, p-1\}$. Since $\{0, \ldots, p-1\}$ is a finite set, but the sequence is infinite, *two distinct elements of the sequence in* (1.10) *must have the same mod p values*. That is, there are positive integers j, m such that

$$z^j = z^{j+m} \pmod{p} \tag{1.11}$$

Since $z \in \{1, \ldots, p-1\}$, we know $z \neq 0 \pmod{p}$. Lemma 1.10(b) ensures that

$$z^j \neq 0 \pmod{p}$$

Define the integer $a = z^j$. The above equality ensures a is not a multiple of p. Substituting

$a = z^j$ in (1.11) gives

$$a = az^m \pmod{p}$$

which means there is an integer k such that

$$az^m - a = kp$$

Thus

$$a(z^m - 1) = kp$$

We have already stated that a is not a multiple of p, so the above equality can only be true if $z^m - 1$ is a multiple of p:

$$z^m - 1 = rp \tag{1.12}$$

for some integer r. Equality (1.12) implies

$$z \cdot z^{m-1} = 1 + rp$$

and so

$$z \cdot z^{m-1} = 1 \pmod{p}$$

Define $s = \mod(z^{m-1}, p)$. Then $s \in \{0, \ldots, p-1\}$ and the above equality implies

$$zs = 1 \pmod{p}$$

which means s is a multiplicative inverse for z and we are done: Axiom 7 holds.

1.8. Exercises

EXERCISE 1.1. *Suppose $\mathbb{F}$ is a field. Let x and y be scalars in $\mathbb{F}$. Prove that*

$$(xy = 0) \iff (x = 0 \text{ or } y = 0)$$

Hint: The reverse direction is an immediate application of Lemma 1.2. For the forward direction, assume $x, y \in \mathbb{F}$ and $xy = 0$, then consider the cases $x = 0$ and $x \neq 0$.

EXERCISE 1.2. *Let $\mathbb{F}$ be a field. Lemma 1.3 ensures $1 \neq 0$ and so Axiom 7 ensures 1^{-1} exists. Prove $1^{-1} = 1$.*

EXERCISE 1.3. *Let $\mathbb{F}$ be a field. Use the axioms to prove that for all $x \in \mathbb{F}$ we have $-x = (-1)x$.*

EXERCISE 1.4. *(Uniqueness of multiplicative inverse) Let $\mathbb{F}$ be a field. Suppose $x \in \mathbb{F}$ and $x \neq 0$. By the field axioms, we know there is a scalar $x^{-1} \in \mathbb{F}$ such that $x \cdot x^{-1} = 1$. We want to show x^{-1}*

is the unique *scalar with this property.*

a) Suppose there is an $a \in \mathbb{F}$ such that $x \cdot a = 1$. Show that $a = x^{-1}$.

b) Find i^{-1} when $\mathbb{F} = \mathbb{C}$ is the field of complex numbers.

c) Find 1^{-1} when $\mathbb{F} = \{0, 1\}$ is the binary field.

d) Find 1^{-1}, 2^{-1}, 3^{-1}, and 4^{-1} when $\mathbb{F} = \{0, 1, 2, 3, 4\}$ is the finite field with mod 5 arithmetic.

EXERCISE 1.5. *Let $\mathbb{F}$ be a field. Fix $x, y \in \mathbb{F}$. Suppose $x \neq 0$ and $y \neq 0$.*

a) Show that $xy \neq 0$.

b) Now that we know $xy \neq 0$, which axiom ensures $(xy)^{-1}$ exists?

c) Prove that $(xy)^{-1} = x^{-1}y^{-1}$.

EXERCISE 1.6. *Let $\mathbb{F}$ be a field.*

a) Define $2 = 1 + 1$. Use the axioms to prove that for all $x \in \mathbb{F}$ we have $x + x = 2x$.

b) Define $3 = 2 + 1$. Use the axioms to prove that for all $x \in \mathbb{F}$ we have $x + x + x = 3x$.

EXERCISE 1.7. *(Additive and multiplicative inverses for $\mathbb{C}$) Consider the complex field $\mathbb{C}$. Let $z = a + bi$.*

a) Find $-z$. That is, give an expression for $-z$ in terms of a, b.

b) Suppose $z \neq 0$, that is, $(a, b) \neq (0, 0)$. Give an expression for z^{-1} in terms of a, b.

EXERCISE 1.8. *(mod 82 arithmetic is not a field) Consider $\mathbb{F} = \{0, 1, \ldots, 81\}$ with mod 82 arithmetic. Show that this does not form a field. Which axiom of a field is violated?*

EXERCISE 1.9. *(Imaginary numbers) Let $\mathbb{C}$ be the set of complex numbers. Note that $i \in \mathbb{C}$ and $i = \sqrt{-1}$. Define $\mathbb{I} \subseteq \mathbb{C}$ as the set of purely imaginary numbers: $\mathbb{I} = \{ix : x \in \mathbb{R}\}$. Maintaining the same definitions of complex addition and multiplication in $\mathbb{C}$, does $\mathbb{I}$ form a field?*

EXERCISE 1.10. *(Additive and multiplicative inverses for $\mathbb{F}_5$) Consider the finite field $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ with mod 5 arithmetic.*

a) Compute $-x$ for all $x \in \mathbb{F}_5$. Compute x^{-1} for all nonzero $x \in \mathbb{F}_5$.

b) Using $\mathbb{F}_5$ with mod 5 arithmetic, solve $3x = 2$. Is there a solution x? Is the solution unique?

EXERCISE 1.11. *(Additive and multiplicative inverses for $\mathbb{F}_{11}$) Consider the finite field $\mathbb{F}_{11} = \{0, 1, 2, \ldots, 10\}$ with mod 11 arithmetic.*

a) Compute $-x$ for all $x \in \mathbb{F}_{11}$.

b) Compute x^{-1} for all nonzero $x \in \mathbb{F}_{11}$.

c) Using $\mathbb{F}_{11}$ with mod 11 arithmetic, solve $8x = 9$. Is there a solution x? Is the solution unique?

EXERCISE 1.12. *(Fractions) Suppose $\mathbb{F}$ is a field. Use the definition $\frac{a}{b} = ab^{-1}$ for $a, b \in \mathbb{F}$, $b \neq 0$ to prove*

a) $\frac{a+b}{c} = \frac{a}{c} + \frac{b}{c}$ for $a, b, c \in \mathbb{F}$, $c \neq 0$.

b) $\frac{1}{(\frac{a}{b})} = \frac{b}{a}$ for $a, b \in \mathbb{F} \setminus \{0\}$.

c) $\frac{a}{b} + \frac{c}{d} = \frac{ad+cb}{bd}$ for $a, b, c, d \in \mathbb{F}$, $b \neq 0, d \neq 0$.

EXERCISE 1.13. *(Modular arithmetic)*

a) Compute $mod(2, 12)$, $mod(-4, 12)$, $mod(34, 12)$, $mod(44, 12)$.

b) Compute $mod(2, 7)$, $mod(-4, 7)$, $mod(34, 7)$, $mod(44, 7)$.

c) Compute $mod(2, 4)$, $mod(8, 4)$, $mod(-8, 4)$, $mod(42, 4)$.

EXERCISE 1.14. *(Once per row and column — Addition table) Suppose* $\mathbb{F} = \{x_1, ..., x_n\}$ *is a finite field with n scalars. Since the scalars are distinct, we have*

$$(i, j \in \{1, ..., n\} \text{ and } (i \neq j)) \implies x_i \neq x_j$$

Addition is defined according to an addition table with n rows and n columns. We want to show that every scalar in $\mathbb{F}$ *appears exactly once on each row and column of the addition table. Fix* $i \in \{1, ..., n\}$. *For simplicity, we first show the result holds for row i.*

a) No duplicates in row i: Suppose there are distinct columns j and k (so $j \neq k$*) that have the same scalar in row i, so:*

$$x_i + x_j = x_i + x_k$$

reach a contradiction using existence of $-x_i$.

b) Every scalar appears on row i: Fix $z \in \mathbb{F}$. *We want to show that z is on row i of the addition table. That is, we want to show there is a scalar* $s \in \mathbb{F}$ *such that* $x_i + s = z$. *Explicitly find such an s by using existence of* $-x_i$.

c) Explain how to show the same result for column i (rather than row i).

EXERCISE 1.15. *(Once per row and column except the first — Multiplication table) As in the previous exercise, suppose* $\mathbb{F} = \{x_1, ..., x_n\}$ *is a finite field with n scalars. Assume* $x_1 = 0$ *and* $x_2 = 1$. *Multiplication is defined by some multiplication table with n rows and n columns. Fix* $i \in \{2, ..., n\}$. *Use the argument structure of the previous exercise to show that each scalar in* $\mathbb{F}$ *appears exactly once on row i. Write a sentence to explain why the same holds for column i.*

EXERCISE 1.16. *(Axioms and mod 4) Let* $S = \{0, 1, 2, 3\}$ *with mod 4 arithmetic. The addition and multiplication tables are:*

+	0	1	2	3
0	0	1	2	3
1	1	2	3	0
2	2	3	0	1
3	3	0	1	2

×	0	1	2	3
0	0	0	0	0
1	0	1	2	3
2	0	2	0	2
3	0	3	2	1

It can be shown that 6 out of the 7 axioms for a field hold. However, exactly one axiom fails. Which axiom fails? Give a counter-example to show it fails.

EXERCISE 1.17. *(The field $\mathbb{F}_4$) We want to prove there is a field of size 4. Define $\mathbb{F}_4 = \{0, 1, a, b\}$ as a set with four distinct elements (including 0 and 1 that are the distinct additive and multiplicative identities). We want to find a multiplication table and an addition table such that $\mathbb{F}_4$ (together with this scalar addition and multiplication) satisfies the axioms for a field. The tables below are partially filled out.*

$+$	0	1	a	b
0	0	1	a	b
1				
a				
b				

$\times$	0	1	a	b
0	0	0	0	0
1	0	1	a	b
a				
b				

a) Prove that (similar to Sudoku), each of the four rows of the addition table is a permutation of $\{0, 1, a, b\}$ (so there are no repetitions in a row). Similarly, each of the four columns is a permutation of $\{0, 1, a, b\}$.

b) Prove that the last three rows of the multiplication table is a permutation of $\{0, 1, a, b\}$. Similar for the columns. We now proceed to part (c).

c) Prove that we cannot have either $b \cdot 0 = 1$, $b \cdot 1 = 1$, $b \cdot b = 1$, and so we must have $ba = 1$.

d) Complete the multiplication and addition tables.

e) How can we check that your definition of scalar addition and multiplication satisfies Axiom 1? Axioms 4, 5, 6, 7?

f) How can we check that Axioms 2 and 3 are satisfied? How can a computer be used?

EXERCISE 1.18. *Define the 28-element set $F = \{0, 1, 2, \ldots, 27\}$ and consider mod 28 arithmetic. Show this is not a field.*

EXERCISE 1.19. *(Computer program for Axioms 2-3) Write a computer program to show that $\mathbb{F}_3$ with mod 3 arithmetic satisfies Axioms 2-3 for a field. You are free to cut, paste, and modify the* MATLAB *program in Section 1.4.3.*

EXERCISE 1.20. *(Square roots in a field) Let $\mathbb{F}$ be a field. We want to show every $b \in \mathbb{F}$ has either 0, 1, or 2 square roots (a square root of b is a scalar x such that $x^2 = b$).*

a) Fix $b \in \mathbb{F}$. Suppose scalars x and y are both square roots of b. Use $x^2 = y^2$ to show that either $x = y$ or $x = -y$. Explain why this means every scalar b has at most two square roots.

b) Give an example of a field and a scalar b that has no square roots.

c) Give an example of a field and a scalar b that has exactly one square root.

d) Consider $\mathbb{F}_7 = \{0, 1, 2, 3, 4, 5, 6\}$ with mod 7 arithmetic. How many square roots does 2 have? How many square roots does 3 have?

EXERCISE 1.21. *(Axioms 2 and 3 for $\mathbb{C}$) Complete the proof that Axioms 2 and 3 hold for the complex numbers $\mathbb{C}$ with the rules of complex addition and multiplication given in Section 1.4.2.*

EXERCISE 1.22. *(The rationals $\mathbb{Q}$ form a field) Assume $\mathbb{R}$, together with standard addition and multiplication of real numbers, forms a field. This problem uses that to prove the subset $\mathbb{Q} \subseteq \mathbb{R}$ of rational numbers also forms a field. Specifically, $\mathbb{Q} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}$.*

a) Verify closure *of addition for rational numbers: The sum of two rationals is rational.*

b) Verify closure for rational multiplication.

c) Verify that 0 and 1 are rational. Verify that for real numbers x we have:

$$(x \in \mathbb{Q}) \implies -x \in \mathbb{Q}$$
$$(x \in \mathbb{Q} \text{ and } x \neq 0) \implies x^{-1} \in \mathbb{Q}$$

d) Explain how our assumption that $\mathbb{R}$ is a field ensures $\mathbb{Q}$ is a field.

EXERCISE 1.23. *(Sum of powers of scalars in $\mathbb{F}$) Let $\mathbb{F}$ be a general field. Fix $x \in \mathbb{F}$ such that $x \neq 1$. For $n \in \mathbb{N}$ define $s = \sum_{i=1}^{n} x^i$ where*

$$x^i = \underbrace{xx\cdots x}_{i \text{ times}}$$

a) Give a nice expression for $(1-x)s$. [Hint: You should get nice cancellations.]

b) Explain why

$$\sum_{i=1}^{n} x^i = (x - x^{n+1})(1-x)^{-1} \quad \forall n \in \mathbb{N} \tag{1.13}$$

EXERCISE 1.24. *(What about infinite sums?) Infinite sums of scalars in a general field $\mathbb{F}$ are undefined. This question explores why.*

a) For $\mathbb{F} = \mathbb{R}$, use $x = 2^{-1}$ in formula (1.13) to obtain an expression for $\sum_{i=1}^{n}(1/2)^i$ for all $n \in \mathbb{N}$. Take $n \to \infty$ to show there is a good definition of $\sum_{i=1}^{\infty}(1/2)^i$. [Strictly speaking, the limit makes sense by implicitly using additional properties of $\mathbb{R}$ that allow distance comparisons and notions of convergence.]

b) Consider $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic. Use $x = 2^{-1}$ in formula (1.13) to obtain an expression for $\sum_{i=1}^{n}(1/2)^i$ for all $n \in \mathbb{N}$. Explain why taking $n \to \infty$ does not make sense in this case of field $\mathbb{F}_3$.

CHAPTER 2.

Matrix operations

Matrices have a central role in linear algebra. They can be used to efficiently represent and process data. In Chapter 6 we show that any linear transformation between two finite dimensional vector spaces can be represented by a matrix.

This chapter develops results for matrices with entries in a general scalar field $\mathbb{F}$. Recall that scalars can be added and multiplied. Also, every field $\mathbb{F}$ contains two special scalars called 0 and 1. Unless otherwise stated, all results in this chapter hold for a general $\mathbb{F}$. Specific examples are shown for the cases $\mathbb{F} = \mathbb{R}$ and for finite fields $\mathbb{F} = \mathbb{F}_p$ for p prime.

2.1. Matrices

Let $\mathbb{F}$ be a scalar field. Fix m, n as positive integers. A *matrix A of size $m \times n$ with entries in $\mathbb{F}$* is a collection of mn scalars arranged in a grid with m rows and n columns:

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

We often write $A = (a_{ij})$, where $a_{ij} \in \mathbb{F}$ denotes the scalar entry in row $i \in \{1, \ldots, m\}$ and column $j \in \{1, \ldots, n\}$. It is sometimes useful to write entry a_{ij} as $a_{i,j}$ (which more clearly distinguishes the two subscripts). The entries of a matrix are also called *components*.

For example, with $\mathbb{F} = \mathbb{R}$, the following is a 3×4 matrix:

$$A = \begin{bmatrix} -1 & 2.4 & 0 & 0 \\ 9 & 1 & 0 & -1.5 \\ 1 & 5 & 7.3 & 5 \end{bmatrix}$$

The scalar entries on the first row of A are:

$$a_{11} = -1, \quad a_{12} = 2.4, \quad a_{13} = 0, \quad a_{14} = 0$$

The scalar entry in the second row and third column is 0, that is, $a_{23} = 0$.

2.1.1. Notation

Let m, n be positive integers. Define the set of all $m \times n$ matrices with entries in a field $\mathbb{F}$ by

$$\mathbb{F}^{m \times n} = \{m \times n \text{ matrices with entries in } \mathbb{F}\}$$

Thus, $A \in \mathbb{F}^{m \times n}$ means A has m rows, n columns, and $A = (a_{ij})$ for some entries $a_{ij} \in \mathbb{F}$ for $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, n\}$. When the field of interest is clear we use the concise notation

$$A \sim m \times n$$

to mean that A has size $m \times n$.

2.1.2. Transpose of a matrix

The *transpose* of matrix A, denoted $A^\top$, is the matrix that turns all rows of A into columns. That is, if $A \sim m \times n$, then $A^\top \sim n \times m$ and $A^\top = (b_{ij})$ where

$$b_{ij} = a_{ji} \quad \forall i \in \{1, \ldots, n\}, j \in \{1, \ldots, m\}$$

For example, assuming $\mathbb{F}$ contains scalars $0, 1, 2$, consider

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix} \implies A^\top = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix} \tag{2.1}$$

For this example (2.1), $A \sim 3 \times 2$ and $A^\top \sim 2 \times 3$. If A is a matrix then the transpose of its transpose is itself:

$$(A^\top)^\top = A$$

2.1.3. Types of matrices

For a given $m \times n$ matrix A we say:

- A is *square* if it has the same number of rows and columns ($m = n$).
- A is *tall* if it has more rows than columns ($m > n$).
- A is *wide* if it has more columns than rows ($m < n$).
- A is a *column vector* if it has only one column ($n = 1$).
- A is a *row vector* if it has only one row ($m = 1$).
- A is *symmetric* if $A^\top = A$.

Note that A is tall if and only if $A^\top$ is wide. Also, A is square if and only if $A^\top$ is square. Matrix A can only be symmetric if it is square. A matrix $A \sim m \times n$ can be viewed as a collection of n columns (each column being a $m \times 1$ column vector). It can also be viewed as a collection of m rows (each row being a $1 \times n$ row vector).

2.1.4. Convention

When we introduce a $m \times n$ matrix A, it is implicitly assumed that m and n are some positive integers. Also, unless otherwise stated:

- When several matrices are treated together in the same problem, all matrices are assumed to take values in some common field $\mathbb{F}$.
- Results in Chapters 2-7 hold for any general field $\mathbb{F}$.

2.2. Diagonal, upper triangular, and the identity matrix

This section considers square matrices. The *diagonal entries* of a $n \times n$ matrix $A = (a_{ij})$ are the entries a_{ii} for $i \in \{1, \ldots, n\}$. The diagonal entries of a 3×3 matrix are boxed below:

$$A = \begin{bmatrix} \boxed{2} & 1 & 0 \\ 4 & \boxed{0} & 1 \\ 1 & 1 & \boxed{1} \end{bmatrix}$$

The diagonal entries partition the matrix into three parts: (i) Entries on the diagonal (entries a_{ii}); (ii) Entries above the diagonal (entries a_{ij} for $i < j$); (iii) Entries below the diagonal (entries a_{ij} for $i > j$).

DEFINITION 2.1. *A square matrix $A = (a_{ij})$ is a* diagonal matrix *if $a_{ij} = 0$ whenever $i \neq j$. That is, all off-diagonal entries are zero.*

DEFINITION 2.2. *A square matrix $A = (a_{ij})$ is* upper triangular *if $a_{ij} = 0$ whenever $i > j$. That is, all entries of A that are below the diagonal are zero.*

DEFINITION 2.3. *A square matrix $A = (a_{ij})$ is* lower triangular *if $a_{ij} = 0$ whenever $i < j$. That is, all entries of A that are above the diagonal are zero.*

The transpose of an upper triangular matrix is a lower triangular matrix. A diagonal matrix is both upper triangular and lower triangular. Also, every diagonal matrix is symmetric. Examples are:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -2 \end{bmatrix} \quad \text{[diagonal]}$$

$$\begin{bmatrix} 9 & 8 & 0 \\ 0 & 1 & 3 \\ 0 & 0 & -2 \end{bmatrix} \quad \text{[upper triangular]}$$

$$\begin{bmatrix} 5 & 7 \\ 0 & 1 \end{bmatrix} \quad \text{[upper triangular]}$$

$$\begin{bmatrix} -2 & 0 & 0 \\ 4 & 1 & 0 \\ 1 & 0 & 7 \end{bmatrix} \quad \text{[lower triangular]}$$

$$\begin{bmatrix} 3 & 7 & -1 \\ 7 & 2 & 0 \\ -1 & 0 & 1 \end{bmatrix} \quad \text{[symmetric]}$$

Fix $n \in \mathbb{N}$. An important $n \times n$ diagonal matrix is the one where all n diagonal entries are 1: This is called the *$n \times n$ identity matrix* and is denoted I. The 2×2 and 3×3 identity matrices are shown below:

$$I = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad \text{[2×2 identity matrix]}$$

$$I = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad \text{[3×3 identity matrix]}$$

Every field has scalars 0 and 1, so the $n \times n$ identity matrix makes sense for every field.

2.3. Matrix addition and scalar multiplication

If A and B are two matrices of the same size, and with entries in a common scalar field, we define the sum $A + B$ as the matrix formed by adding the individual entries:

$$A + B = (a_{ij} + b_{ij})$$

For example, with $\mathbb{F} = \mathbb{R}$ we have

$$\begin{bmatrix} 2 & 1 \\ -1 & 8 \\ 0 & 4 \end{bmatrix} + \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 0 & -1 \end{bmatrix} = \begin{bmatrix} 3 & 2 \\ -1 & 9 \\ 0 & 3 \end{bmatrix}$$

In the binary field $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic (so $1 + 1 = 0$) we have

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

Matrix addition is commutative and associative: Assuming matrices A, B, C have the same size and have entries in a common field $\mathbb{F}$, we have

$$A + B = B + A \tag{2.2}$$

$$A + (B + C) = (A + B) + C \tag{2.3}$$

This follows from the commutativity and associativity axioms of scalar addition. Formulas (2.2)-(2.3) imply that we can drop the parentheses when summing three matrices:

$$A + B + C = (A + B) + C$$

By induction, it follows that n matrices $A_1, \ldots, A_n$ of the same size can be summed in any order and with any placement of the parentheses:

$$\sum_{i=1}^{n} A_i = A_1 + A_2 + \cdots + A_n = (A_1 + (A_2 + \cdots + A_n))$$

Exercise 2.2 shows that

$$\left(\sum_{i=1}^{n} A_i\right)^{\top} = \sum_{i=1}^{n} A_i^{\top}$$

The multiplication of a scalar c and a matrix A, denoted cA, is defined as the matrix of

the same size as A formed by multiplying each entry of A by c:

$$A = (a_{ij}) \implies cA = (ca_{ij})$$

For example, with $\mathbb{F} = \mathbb{R}$

$$5\begin{bmatrix} 0 & 1 \\ 2 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 5 \\ 10 & 10 \\ 15 & 20 \end{bmatrix} \tag{2.4}$$

For matrices A, B of the same size, we define:

$$A - B = A + (-1)B$$

Scalar multiplication distributes: If matrices A, B have the same size and $c, d \in \mathbb{F}$ then

$$c(A + B) = cA + cB \quad , \quad (c + d)A = cA + dA$$

These equalities are proven in Exercise 2.36. As specified in Section 2.1.4, these operations implicitly assume a common field $\mathbb{F}$. As an example for matrices with entries in $\mathbb{R}$,

$$5\left(\begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix} + \begin{bmatrix} 4 & 4 \\ 1 & 1 \end{bmatrix}\right) = 5\begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix} + 5\begin{bmatrix} 4 & 4 \\ 1 & 1 \end{bmatrix}$$

$$(4 + -2)\begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix} = 4\begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix} + (-2)\begin{bmatrix} 1 & 0 \\ 2 & 3 \end{bmatrix}$$

2.4. Matrix multiplication

A n-tuple is an ordered list of the form $(x_1, ..., x_n)$. Define $\mathbb{F}^n$ as the set of all n-tuples of scalars in $\mathbb{F}$:

$$\mathbb{F}^n = \{(x_1, ..., x_n) : x_i \in \mathbb{F} \quad \forall i \in \{1, \ldots, n\}\}$$

$$(x \in \mathbb{F}^n) \implies (x = (x_1, ..., x_n) \text{ for some scalars } x_i \in \mathbb{F} \text{ for } i \in \{1, ..., n\})$$

We shall call elements of $\mathbb{F}^n$ *vectors*. We shall call $\mathbb{F}^n$ a *vector space*. [More general vector spaces are developed in Chapter 6, including vectors that are not n-tuples.] We do not need to assign a horizontal or vertical orientation to a vector $x \in \mathbb{F}^n$. However, for purposes of matrix arithmetic, it is standard to view $x \in \mathbb{F}^n$ as a column vector (a $n \times 1$ matrix), so its transpose $x^\top$ is a row vector (a $1 \times n$ matrix).

2.4.1. Dot product

Define the *dot product* between two vectors $x, y \in \mathbb{F}^n$, where $x = (x_1, ..., x_n)$ and $y = (y_1, ..., y_n)$, as

$$\text{dot}(x, y) = \sum_{i=1}^{n} x_i y_i$$

Thus, $\text{dot}(x, y)$ takes two vectors $x, y \in \mathbb{F}^n$ and outputs a scalar in $\mathbb{F}$. We can view $\text{dot}(x, y)$ as software that does not care if its input vectors x, y have a row or column orientation, provided: (i) x and y have the same number of components; (ii) the components are scalars in a common field. Alternative notation for the dot product uses a single dot:

$$x \cdot y = \text{dot}(x, y) = \sum_{i=1}^{n} x_i y_i$$

2.4.2. Matrix multiplication via the dot product

Given two matrices A and B with entries in a common field, we define the matrix multiplication AB whenever the number of columns of A is the same as the number of rows of B. Specifically, fix positive integers m, n, k and suppose

$$A \sim m \times n$$
$$B \sim n \times k$$

Then AB is defined as the $m \times k$ matrix $G = (g_{ij})$ where

$$\begin{aligned} g_{ij} &= \text{dot}(\text{row } i \text{ of } A, \text{column } j \text{ of } B) \\ &= \sum_{u=1}^{n} a_{iu} b_{uj} \quad \forall i \in \{1, ..., m\}, j \in \{1, ..., k\} \end{aligned}$$

For a general matrix $D = (d_{ij})$ we can define notation $(D)_{ij}$ to represent the (i, j) entry

$$(D)_{ij} = d_{ij}$$

With this notation, the above formula for entry (i, j) of the product matrix AB is:

$$\boxed{(AB)_{ij} = \sum_{u=1}^{n} a_{iu} b_{uj} \quad \forall i \in \{1, ..., m\}, j \in \{1, ..., k\}} \tag{2.5}$$

By viewing vectors $x, y \in \mathbb{F}^n$ as column vectors, the dot product can be expressed as

matrix multiplication of the $1 \times n$ matrix $x^\top$ and the $n \times 1$ matrix y:

$$x^\top y = \mathrm{dot}(x, y) = \sum_{i=1}^{n} x_i y_i$$

For example

- With $\mathbb{F} = \mathbb{R}$ we have
$$\begin{bmatrix} 2 & 0 & 1 \\ -1 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 3 & 1 \\ -1 & -1 \end{bmatrix}$$
- In the binary field $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic we have
$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}$$
- With $\mathbb{F}_3 = \{0, 1, 2\}$ and mod 3 arithmetic we have
$$\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}$$

2.4.3. Associative and distributive properties of matrices

Matrix multiplication is associative and distributive: Given any matrices A, B, C, D with entries in some field $\mathbb{F}$, and with sizes that make sense for matrix arithmetic, we have

$$(AB)C = A(BC) \tag{2.6}$$

$$(B + D)C = BC + DC \quad , \quad A(B + D) = AB + AD \tag{2.7}$$

We now prove (2.6). [Exercise 2.12 proves (2.7).] Let A, B, C be matrices such that

$$A = (a_{ij}) \sim m \times n$$
$$B = (b_{ij}) \sim n \times k$$
$$C = (c_{ij}) \sim k \times r$$

Define

$$E = AB \quad , \quad F = BC$$

To prove (2.6), it suffices to prove $EC = AF$. For this we show

$$(EC)_{ij} = (AF)_{ij}$$

for all entries (i, j), where $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, r\}$. To this end, write $E = (e_{wz})$ and $F = (f_{wz})$, where (w, z) is a dummy variable for a generic entry of the matrices. Fix (i, j). Then (using formula (2.5) for matrix products):

$$(EC)_{ij} = \sum_u e_{iu} c_{uj} \tag{2.8}$$

where

$$\begin{aligned} e_{iu} &= (AB)_{iu} \\ &= \sum_v a_{iv} b_{vu} \end{aligned}$$

Substituting this into (2.8) gives

$$\begin{aligned} (EC)_{ij} &= \sum_u \left(\sum_v a_{iv} b_{vu} \right) c_{uj} \\ &= \sum_u \left(\sum_v a_{iv} b_{vu} c_{uj} \right) \end{aligned}$$

where scalar c_{uj} has been distributed into the inner sum. Removing the parentheses gives

$$(EC)_{ij} = \sum_u \sum_v a_{iv} b_{vu} c_{uj} \tag{2.9}$$

On the other hand we have

$$(AF)_{ij} = \sum_v a_{iv} f_{vj} \tag{2.10}$$

where

$$\begin{aligned} f_{vj} &= (BC)_{vj} \\ &= \sum_u b_{vu} c_{uj} \end{aligned}$$

Substituting this into (2.10) gives

$$\begin{aligned} (AF)_{ij} &= \sum_v a_{iv} \left(\sum_u b_{vu} c_{uj} \right) \\ &= \sum_v \left(\sum_u a_{iv} b_{vu} c_{uj} \right) \end{aligned}$$

Removing the parentheses gives

$$\begin{aligned}(AF)_{ij} &= \sum_v \sum_u a_{iv} b_{vu} c_{uj} \\ &= \sum_u \sum_v a_{iv} b_{vu} c_{uj}\end{aligned} \tag{2.11}$$

where we have used the *sum switching rule* of (1.2) from Chapter 1.1. Comparing (2.11) and (2.9) shows that $(EC)_{ij} = (AF)_{ij}$. Since this holds for all entries (i, j), we know $EC = AF$ and we are done.

[Note: When writing sums it is usually encouraged to give complete summation limits, such as $\sum_{u=1}^{k}$ instead of $\sum_u$. However, for the derivation of $(AB)C = A(BC)$ it is difficult to specify each summation limit. We have opted for $\sum_u$ notation in this case.]

2.4.4. Multiplying many matrices

Since matrix multiplication is associative, the multiplication $(AB)C$ can be written without parentheses as ABC. The multiplication ABC makes sense whenever the sizes satisfy $A \sim m \times n$, $B \sim n \times k$, $C \sim k \times r$ for some positive integers m, n, k, r. For a collection of c matrices $A_1, \ldots, A_c$ with sizes that make sense for multiplication, their product (in that order) is written $A_1 A_2 \cdots A_c$. The result of $A_1 A_2 \cdots A_c$ is a new matrix with a number of rows equal to the number of rows of A_1 and a number of columns equal to the number of columns of A_c. These size relationships can be expressed for multiplication chains of two or three matrices as

$$\underbrace{A}_{(m\times n)}\ \underbrace{B}_{(n\times k)} = \underbrace{AB}_{(m\times k)}$$

$$\underbrace{A}_{(m\times n)}\ \underbrace{B}_{(n\times k)}\ \underbrace{C}_{(k\times r)} = \underbrace{ABC}_{(m\times r)}$$

In general the size of $A_1 A_2 \cdots A_c$ is

$$\underbrace{A_1}_{(m_1\times m_2)}\ \underbrace{A_2}_{(m_2\times m_3)}\ \underbrace{A_3}_{(m_3\times m_4)}\ \cdots\ \underbrace{A_c}_{(m_c\times m_{c+1})} = \underbrace{A_1 A_2 \cdots A_c}_{(m_1\times m_{c+1})}$$

If A is a square matrix then it can be multiplied by itself as many times as we like. Define

$$A^k = \underbrace{AA\cdots A}_{k \text{ times}} \qquad \forall k \in \mathbb{N}$$

2.4.5. Distributing across many matrices

The formula $A(B+D) = AB + AD$ from (2.7) can be extended. Let $B_1, \ldots, B_s$ be matrices of the same size. Let A be a matrix with the same number of columns as the number of rows of B_1. The entries of all matrices are scalars in some common field $\mathbb{F}$. Then

$$A\left(\sum_{i=1}^{s} B_i\right) = \sum_{i=1}^{s} AB_i \tag{2.12}$$

This is formally proven by induction in Exercise 2.39. From this, we obtain the following easy extension: If $\theta_1, \ldots, \theta_s$ are any scalars in $\mathbb{F}$ then

$$A\left(\sum_{i=1}^{s} \theta_i B_i\right) = \sum_{i=1}^{s} \theta_i AB_i \tag{2.13}$$

This is an "easy extension" because it follows by defining matrices $\tilde{B}_i = \theta_i B_i$.

2.4.6. Is matrix multiplication commutative?

Matrix multiplication is not commutative. For example, if $A \sim 4 \times 8$ and $B \sim 8 \times 2$, then AB makes sense but BA is not defined due to size mismatch. If $A \sim m \times n$ and $B \sim n \times m$, where $m \neq n$, then both AB and BA make sense, but they are two square matrices of different sizes (so they *cannot* be equal). Even if A and B are square and of the same size, we might have $AB \neq BA$. For example, consider $\mathbb{F} = \mathbb{R}$ and

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 1 & 0 \\ 2 & 0 \end{bmatrix} \tag{2.14}$$

Then

$$AB = \begin{bmatrix} 3 & 0 \\ 2 & 0 \end{bmatrix} \quad , \quad BA = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$$

If the same A and B matrices in (2.14) are treated in $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic (so $3 = 0$) then

$$AB = \begin{bmatrix} 0 & 0 \\ 2 & 0 \end{bmatrix} \quad , \quad BA = \begin{bmatrix} 1 & 1 \\ 2 & 2 \end{bmatrix}$$

where the entry of AB on the first row and column is 0 because

$$(1)(1) + (1)(2) = 3 = 0 \pmod 3$$

Since the entries are integers, the resulting product mod 3 can be obtained by taking the

usual matrix product (using field $\mathbb{R}$) and then converting each entry to its corresponding value in the set $\{0, 1, 2\}$ under the mod 3 operation (see Lemma 1.8).

It is possible for two square matrices A and B of the same size to satisfy $AB = BA$. Such matrices are said to *commute*. It is clear that every square matrix commutes with itself because $AA = AA$. Also, any two $n \times n$ diagonal matrices commute. For example,

$$\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix}\begin{bmatrix} b_{11} & 0 & 0 \\ 0 & b_{22} & 0 \\ 0 & 0 & b_{33} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & 0 & 0 \\ 0 & a_{22}b_{22} & 0 \\ 0 & 0 & a_{33}b_{33} \end{bmatrix}$$

$$\begin{bmatrix} b_{11} & 0 & 0 \\ 0 & b_{22} & 0 \\ 0 & 0 & b_{33} \end{bmatrix}\begin{bmatrix} a_{11} & 0 & 0 \\ 0 & a_{22} & 0 \\ 0 & 0 & a_{33} \end{bmatrix} = \begin{bmatrix} a_{11}b_{11} & 0 & 0 \\ 0 & a_{22}b_{22} & 0 \\ 0 & 0 & a_{33}b_{33} \end{bmatrix}$$

Exercise 2.51 shows that if A is any square matrix, I is the identity matrix of the same size, and c, d are two scalars, then the matrices $(A - cI)$ and $(A - dI)$ commute.

2.5. Data crunching with matrices

In many applications we want to take data in the form of a vector $x \in \mathbb{F}^n$ and crunch it to another vector $y \in \mathbb{F}^m$ in some structured way. (For example, we may have $n = 1000$ and $m = 2$.) This can be done by left-multiplying x by some fixed $m \times n$ matrix $A = (a_{ij})$:

$$y = Ax$$

Formally this input-output response is specified by the function $T : \mathbb{F}^n \to \mathbb{F}^m$ defined by

$$T(x) = Ax \quad \forall x \in \mathbb{F}^n$$

Such a function is a *linear transformation* (linear transformations are studied in detail in Chapter 6). Not all functions from $\mathbb{F}^n$ to $\mathbb{F}^m$ are linear transformations. However, the matrix $A = (a_{ij})$ has many coefficients a_{ij} to wiggle, so desired input-output behavior can often be captured by some choice of these coefficients. A more general input-output response can use a function $R : \mathbb{F}^n \to \mathbb{F}^m$ defined by

$$R(x) = Ax + b \quad \forall x \in \mathbb{F}^n$$

where $A = (a_{ij})$ is some fixed $m \times n$ matrix and b is some fixed vector in $\mathbb{F}^n$. Such a function R is called an *affine transformation*.

Another common scenario is to have an unknown system that takes as input a vector

$x \in \mathbb{R}^n$ and outputs a scalar $y \in \mathbb{R}$. The input-output behavior of the system acts according to some unknown model, possibly a model that has nonlinearities and that adds random noise to the output. We want a *predictor* that takes a potential input vector $x \in \mathbb{R}^n$ and outputs a *prediction* $\hat{y} \in \mathbb{R}$ that is hopefully close to the actual output y. Even when the actual input-output of the system has nonlinearities and randomness, accurate predictors can often be built using the theory of *linear regression*. Linear regression uses the theory of linear transformations to build a predictor (using matrices and matrix transposes) from observed *data* about prior input-output responses. This topic is studied in Chapter 9.

2.6. Transpose of a product of matrices

LEMMA 2.1. *If $A \sim m \times n$ and $B \sim n \times k$, then*

$$(AB)^\top = B^\top A^\top$$

PROOF. Fix $i \in \{1, \ldots, m\}$ and $j \in \{1, \ldots, k\}$. Let d_{ij} be the (i, j) entry of AB. Then d_{ij} is the dot product of row i of A and column j of B. But row i of A is the same as column i of $A^\top$, column j of B is the same as row j of $B^\top$. So d_{ij} is equal to the (j, i) entry of $B^\top A^\top$. □

From the above lemma we obtain

$$(ABC)^\top = C^\top (AB)^\top = C^\top B^\top A^\top$$

assuming the sizes make sense for the multiplication ABC. For k matrices (with appropriate sizes so the multiplication $A_1 A_2 \cdots A_k$ makes sense), the formula becomes

$$(A_1 A_2 \cdots A_{k-1} A_k)^\top = A_k^\top A_{k-1}^\top \cdots A_2^\top A_1^\top \tag{2.15}$$

An important consequence of Lemma 2.1 is this: If A is a matrix of any size $m \times n$, then (see Exercise 2.27):

- $A^\top A$ is a symmetric $n \times n$ matrix.
- $AA^\top$ is a symmetric $m \times m$ matrix.

The fact that both $A^\top A$ and $AA^\top$ are symmetric is useful for the theory of *singular values* (treated in Chapter 16).

2.7. Trace of a matrix

If A is a square matrix with entries in some field, we define its *trace*, written $\text{trace}(A)$, as the sum of its diagonal entries:

$$\Big(A = (a_{ij}) \sim n \times n\Big) \implies \text{trace}(A) = \sum_{i=1}^{n} a_{ii}$$

For example,

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \implies \text{trace}(A) = a_{11} + a_{22} + a_{33}$$

It is clear that if A and B are square matrices of the same size and with entries in $\mathbb{F}$, and if θ is a scalar in $\mathbb{F}$, then

$$\begin{aligned} &\text{trace}(A + B) = \text{trace}(A) + \text{trace}(B) \\ &\text{trace}(\theta A) = \theta \text{trace}(A) \\ &\text{trace}(A) = \text{trace}(A^\top) \end{aligned}$$

An interesting property is that if $C \sim m \times n$ and $D \sim n \times m$, where m and n are any positive integers, then both CD and DC are square (with possibly different sizes) and

(2.16) $$\boxed{\text{trace}(CD) = \text{trace}(DC)}$$

This is formally proven in Exercise 2.16. For example, using field $\mathbb{F} = \mathbb{R}$, consider

$$C = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 4 & 9 \end{bmatrix} \quad , \quad D = \begin{bmatrix} -1 & 0 \\ 11 & 4 \\ 2 & 2 \end{bmatrix}$$

$$CD = \begin{bmatrix} 27 & 14 \\ 62 & 34 \end{bmatrix} \quad , \quad DC = \begin{bmatrix} -1 & -2 & -3 \\ 11 & 38 & 69 \\ 2 & 12 & 24 \end{bmatrix}$$

$$\text{trace}(CD) = 27 + 34 = 61 \quad , \quad \text{trace}(DC) = -1 + 38 + 24 = 61$$

2.8. Structural properties of matrix multiplication

2.8.1. Multiplying a matrix and a vector

Let A be a $m \times n$ matrix (with entries in $\mathbb{F}$). We can think of A in terms of its n columns. We can also think of A in terms of its m rows. Specifically:

$$A = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} - & r_1^\top & - \\ - & r_2^\top & - \\ \vdots & \vdots & \vdots \\ - & r_m^\top & - \end{bmatrix}$$

- For $j \in \{1, \ldots, n\}$, let $c_j \in \mathbb{F}^m$ be the jth column of A (viewing c_j as a column vector).
- For $i \in \{1, \ldots, m\}$, let $r_i^\top \in \mathbb{F}^n$ be the ith row of A (viewing $r_i^\top$ as a row vector).

Let $x = (x_1, \ldots, x_n)$ be a column vector in $\mathbb{F}^n$. The multiplication Ax is defined by viewing x as a $n \times 1$ matrix, so the size of Ax is

$$\underbrace{(m \times n)}_{\text{size}(A)} \underbrace{(n \times 1)}_{\text{size}(x)} = \underbrace{m \times 1}_{\text{size}(Ax)}$$

In particular, the rules of matrix multiplication ensure that Ax is a column vector that is a weighted sum of the columns of A:

$$Ax = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = x_1 \begin{bmatrix} | \\ c_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ c_2 \\ | \end{bmatrix} + \cdots + x_n \begin{bmatrix} | \\ c_n \\ | \end{bmatrix}$$

In particular

$$\boxed{Ax = \sum_{j=1}^{n} x_j c_j}$$

The weighted sum on the right-hand-side of the above equality is called a *linear combination of the columns of* A. The "weights" are the scalars $x_1, \ldots, x_n \in \mathbb{F}$ and the columns of A are $c_1, \ldots, c_n \in \mathbb{F}^m$. The term *linear combination* is formally defined for vectors in $\mathbb{F}^m$ in Chapter 3.1 and for vectors in a general vector space V in Chapter 6.2.

Let $y = (y_1, \ldots, y_m)$ be a column vector in $\mathbb{F}^m$. The multiplication $y^\top A$ is defined by

viewing $y^\top$ as a $1 \times m$ matrix, so the size of $y^\top A$ is

$$\underbrace{(1 \times m)}_{\text{size}(y^\top)} \underbrace{(m \times n)}_{\text{size}(A)} = \underbrace{1 \times n}_{\text{size}(y^\top A)}$$

The rules of matrix multiplication ensure that $y^\top A$ is a row vector that is a weighted sum of the rows of A:

$$y^\top A = \begin{bmatrix} y_1 & y_2 & \cdots & y_m \end{bmatrix} \begin{bmatrix} — & r_1^\top & — \\ — & r_2^\top & — \\ \vdots & \vdots & \vdots \\ — & r_m^\top & — \end{bmatrix} = y_1 \begin{bmatrix} — & r_1^\top & — \end{bmatrix} + \cdots + y_m \begin{bmatrix} — & r_m^\top & — \end{bmatrix}$$

In particular

$$y^\top A = \sum_{i=1}^{m} y_i r_i^\top$$

2.8.2. Multiplying two matrices

Let $A \sim m \times k$ and $B \sim k \times n$. Define $C = AB$, so $C \sim m \times n$ and

$$\underbrace{\begin{bmatrix} — & u_1^\top & — \\ — & u_2^\top & — \\ \vdots & \vdots & \vdots \\ — & u_m^\top & — \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & \cdots & | \end{bmatrix}}_{B} = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}}_{C} \tag{2.17}$$

- $u_1^\top, \ldots, u_m^\top$ denotes the m rows of A.
- $v_1, \ldots, v_n$ denotes the n columns of B.
- $c_1, \ldots, c_n$ denotes the n columns of C.

The rules of matrix multiplication ensure that for each $j \in \{1, \ldots, n\}$, the jth column of AB is equal to Av_j:

$$AB = A \underbrace{\begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & \cdots & | \end{bmatrix}}_{B} = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ Av_1 & Av_2 & \cdots & Av_n \\ | & | & \cdots & | \end{bmatrix}}_{C}$$

and so

$$(j\text{th column of } C) = c_j = Av_j \quad \forall j \in \{1, \ldots, n\}$$

Observe that the jth column of AB depends only on the matrix A and on the jth column of matrix B (the other columns of B have no impact).

On the other hand, the rules of matrix multiplication ensure that for each $i \in \{1, \ldots, m\}$, the ith row of AB is equal to $u_i^\top B$:

$$AB = \underbrace{\begin{bmatrix} — & u_1^\top & — \\ — & u_2^\top & — \\ \vdots & \vdots & \vdots \\ — & u_m^\top & — \end{bmatrix}}_{A} B = \underbrace{\begin{bmatrix} — & u_1^\top B & — \\ — & u_2^\top B & — \\ \vdots & \vdots & \vdots \\ — & u_m^\top B & — \end{bmatrix}}_{C}$$

Let $r_1^\top, \ldots, r_m^\top$ denote the rows of C. Then

$$(i\text{th row of } C) = r_i^\top = u_i^\top B \quad \forall i \in \{1, \ldots, m\}$$

Observe that the ith row of AB depends only on the matrix B and on the ith row of matrix A (the other rows of A have no impact).

2.8.3. Block multiplication of matrices

It is sometimes useful to view a matrix as being composed of smaller blocks of matrices. Suppose $A \sim (m + r) \times (n + k)$ and $B \sim (n + k) \times (s + u)$

$$A = \begin{bmatrix} C_{m\times n} & D_{m\times k} \\ E_{r\times n} & F_{r\times k} \end{bmatrix} \quad , \quad B = \begin{bmatrix} J_{n\times s} & K_{n\times u} \\ L_{k\times s} & M_{k\times u} \end{bmatrix}$$

where $C_{m\times n}$ is a submatrix of size $m \times n$, $D_{m\times k}$ is a submatrix of size $m \times k$, and so on. Here, m, r, n, k, s, u can be any positive integers, though the submatrix sizes are organized so the multiplication AB makes sense. In this case, the laws of matrix multiplication ensure

$$AB = \begin{bmatrix} C_{m\times n} & D_{m\times k} \\ E_{r\times n} & F_{r\times k} \end{bmatrix} \begin{bmatrix} J_{n\times s} & K_{n\times u} \\ L_{k\times s} & M_{k\times u} \end{bmatrix} = \begin{bmatrix} CJ + DL & CK + DM \\ EJ + FL & EK + FM \end{bmatrix}$$

where the four blocks of the final matrix are

$$CJ + DL = C_{m\times n}J_{n\times s} + D_{m\times k}L_{k\times s} \quad [\text{size } m \times s]$$

$$CK + DM = C_{m\times n}K_{n\times u} + D_{m\times k}M_{k\times u} \quad [\text{size } m \times u]$$
$$EJ + FL = E_{r\times n}J_{n\times s} + F_{r\times k}L_{k\times s} \quad [\text{size } r \times s]$$
$$EK + FM = E_{r\times n}K_{n\times u} + F_{r\times k}M_{k\times u} \quad [\text{size } r \times u]$$

A simpler version of this is when one matrix multiplies another matrix composed of two blocks:

$$C\begin{bmatrix} J & K \end{bmatrix} = \begin{bmatrix} CJ & CK \end{bmatrix}$$
$$\begin{bmatrix} C \\ E \end{bmatrix} J = \begin{bmatrix} CJ \\ EJ \end{bmatrix}$$

where the number of columns of C is assumed to be the same as the number of rows of both J and K; matrices C and E have the same number of columns.

Fix $q \in \mathbb{N}$ with $q \geq 2$. When there are many block matrices $J_1, \ldots, J_q$, each having the same number of rows as the number of columns in a matrix C:

$$C\begin{bmatrix} J_1 & J_2 & \cdots & J_q \end{bmatrix} = \begin{bmatrix} CJ_1 & CJ_2 & \cdots & CJ_q \end{bmatrix}$$

When there are many block matrices $C_1, \ldots, C_q$, each having the same number of columns as the number of rows of a matrix J:

$$\begin{bmatrix} C_1 \\ C_2 \\ \vdots \\ C_q \end{bmatrix} J = \begin{bmatrix} C_1J \\ C_2J \\ \vdots \\ C_qJ \end{bmatrix}$$

2.9. Multiplying with the identity matrix

Let $e_1, e_2, \ldots, e_n$ denote the columns of the $n \times n$ identity matrix I:

$$I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & \cdots & 1 \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \\ e_1 & e_2 & \cdots & e_n \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} — & e_1^\top & — \\ — & e_2^\top & — \\ \vdots & \vdots & \vdots \\ — & e_n^\top & — \end{bmatrix}$$

For each $i \in \{1, \ldots, n\}$, e_i is a vector in $\mathbb{F}^n$ that has a 1 in its ith entry, and is zero in all other entries. In the above equation e_i is viewed as a column vector. The matrix I is symmetric so $I^\top = I$ and $e_1^\top, e_2^\top, \ldots, e_n^\top$ form the rows of I. The collection of vectors $\{e_1, \ldots, e_n\}$ is called the *standard basis for* $\mathbb{F}^n$ (the term *basis* is formally defined in Chapter 3.9).

The matrix I is called the *identity matrix* because multiplying any other matrix with I, either on the right or on the left, does not change that other matrix (assume sizes are appropriate). Indeed, consider matrices $A \sim m \times n$ and $B \sim n \times k$ with entries in $\mathbb{F}$. If I is the $n \times n$ identity matrix then

$$AI = A \quad , \quad IB = B$$

Specifically,

$$AI = A \underbrace{\begin{bmatrix} | & | & \cdots & | \\ e_1 & e_2 & \cdots & e_n \\ | & | & \cdots & | \end{bmatrix}}_{I} = \begin{bmatrix} | & | & \cdots & | \\ Ae_1 & Ae_2 & \cdots & Ae_n \\ | & | & \cdots & | \end{bmatrix} = A \sim (m \times n)$$

$$IB = \underbrace{\begin{bmatrix} - & e_1^\top & - \\ - & e_2^\top & - \\ \vdots & \vdots & \vdots \\ - & e_n^\top & - \end{bmatrix}}_{I} B = \begin{bmatrix} - & e_1^\top B & - \\ - & e_2^\top B & - \\ \vdots & \vdots & \vdots \\ - & e_n^\top B & - \end{bmatrix} = B \sim (n \times k)$$

- Ae_j is the jth column of matrix A (for each $j \in \{1, \ldots, n\}$).
- $e_i^\top B$ is the ith row of matrix B (for each $i \in \{1, \ldots, n\}$).

For example, with $\mathbb{F} = \mathbb{R}$ we have

$$\begin{bmatrix} 4 & 0 & 1 & 3 \\ 3 & 3 & 3 & 3 \\ 8 & 9 & -2 & 77 \end{bmatrix} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{I} = \begin{bmatrix} 4 & 0 & 1 & 3 \\ 3 & 3 & 3 & 3 \\ 8 & 9 & -2 & 77 \end{bmatrix}$$

For any field we have

$$\underbrace{\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}_{I} \begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Multiplying I by itself does not change it, so

$$I^k = I \quad \forall k \in \{1, 2, 3, \ldots\}$$

2.10. Invertible matrices

DEFINITION 2.4. *A square matrix A is* invertible *if there is some square matrix B of the same size such that*

$$AB = BA = I$$

where I is the identity matrix with the same size as A. If such a matrix B exists, it must be unique and is called the inverse *of matrix A. We let A^{-1} denote this inverse (so $A^{-1} = B$).*

The next lemma justifies uniqueness of the inverse.[1]

LEMMA 2.2. *(Uniqueness of inverse matrix) Suppose A is square, and there are two matrices B and C that both satisfy the requirements for an inverse of A, so*

$$AB = BA = I$$
$$AC = CA = I$$

Then $B = C$. And of course A is invertible and $A^{-1} = B$.

PROOF. Since multiplication by I gives the same matrix back, we have

$$\begin{aligned} C &= CI \\ &\overset{(a)}{=} C(AB) \\ &= (CA)B \\ &\overset{(b)}{=} IB \\ &= B \end{aligned}$$

where (a) holds because we are told $AB = I$; (b) holds because we are told $CA = I$. □

The above proof only uses the assumptions $AB = I$ and $CA = I$. This leads to the following more general lemma.

LEMMA 2.3. *Suppose A is square, and there are two square matrices B and C that satisfy*

$$AB = I \quad , \quad CA = I$$

Then $B = C$, A is invertible, and $A^{-1} = B$.

PROOF. The proof is similar and is left as an exercise (Exercise 2.34). □

[1] For general nonempty sets U, V, a function $f : U \to V$ is said to be *invertible* if there is a function $g : V \to U$ such that $f(g(v)) = v$ for all $v \in V$ and $g(f(u)) = u$ for all $u \in U$. The definition of *invertible matrix* A corresponds to the functions $f : \mathbb{F}^n \to \mathbb{F}^n$ and $g : \mathbb{F}^n \to \mathbb{F}^n$ defined by $f(x) = Ax$ and $g(y) = A^{-1}y$ for $x, y \in \mathbb{F}^n$.

Lemma 2.3 can be used to establish a more general result: If A and B are square matrices that satisfy $AB = I$, then $AB = BA = I$, so both A and B are invertible and they are inverses of each other. The proof of this is postponed until Chapter 5 (it uses a *rank-nullity* result developed there). Here is an intuitive lemma, its proof is in Exercise 2.19:

LEMMA 2.4. *If a matrix A is invertible, then its inverse A^{-1} is also invertible and*

$$(A^{-1})^{-1} = A$$

2.10.1. Multiplying upper triangular matrices

It can be shown that the product of two square and upper triangular matrices is again upper triangular (Exercise 2.21). So, if U_1, U_2 are upper triangulars of the same size then

$$U_1 U_2 = U_3$$

where U_3 is also square and upper triangular.

What about lower triangular matrices? The above result about upper triangulars can be used to prove that the multiplication of two lower triangulars is again lower triangular: Suppose L_1 and L_2 are two square and lower triangular matrices; let C be their output when multiplied:

$$L_1 L_2 = C$$

By taking transposes of this equation (using Lemma 2.1), it follows that $C^\top$ is the product of two upper triangulars, so $C^\top$ is upper triangular, so C is lower triangular.

If a matrix A is both invertible and upper triangular, it can be shown that its inverse A^{-1} is also upper triangular (see Exercise 2.20; see also discussion in Chapter 4.11.2).

2.10.2. Nonsquare matrices can multiply to I

By definition, only square matrices can be invertible. However, it is possible for two nonsquare matrices to multiply to a (square) identity matrix:

- In any field $\mathbb{F}$:

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

- In the field $\mathbb{R}$:

$$\begin{bmatrix} \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{3}} \\ 0 & \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{2}} \\ \frac{1}{\sqrt{3}} & \frac{1}{\sqrt{2}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Nonsquare matrices that multiply to I are related to the concept of a *pseudoinverse*, defined in Appendix B. See also Exercises 2.45 and 2.46.

2.10.3. Terminology

- A square and noninvertible matrix is also called a *singular matrix*.
- An invertible matrix is also called a *nonsingular matrix*.

2.10.4. Solving $Ax = b$ when A is invertible

Suppose we are given a square matrix A of size $n \times n$ with entries in a field $\mathbb{F}$, and a column vector $b \in \mathbb{F}^n$. We are interested in finding vectors $x \in \mathbb{F}^n$ that satisfy the equation

$$Ax = b$$

Important questions are:

- Does there exist a solution x?
- If a solution exists, is it unique? Can there be more than one solution?
- Does the equation $Ax = b$ have a solution for all $b \in \mathbb{F}^n$? If not, for which b can it be solved?

These questions are easy to answer when A is invertible.

LEMMA 2.5. *Suppose A is an invertible $n \times n$ matrix with entries in some field $\mathbb{F}$. Then*
a) For all $b \in \mathbb{F}^n$, the equation $Ax = b$ has one and only one solution, namely

$$x = A^{-1}b$$

b) The only solution to the equation $Ax = 0$ is $x = 0$ (where $0 = (0, \ldots, 0)$ denotes the all-zero vector in $\mathbb{F}^n$). The equation $Ax = 0$ is called the homogeneous equation *and the solution $x = 0$ is called the* trivial solution.

PROOF. To prove (a), fix $b \in \mathbb{F}^n$. Define $x = A^{-1}b$. Then

$$Ax = AA^{-1}b = Ib = b$$

and so x solves the equation $Ax = b$. To show that $A^{-1}b$ is the *only* solution, suppose z is some other vector that satisfies

$$Az = b$$

Multiplying both sides of the above equation by A^{-1} gives

$$\underbrace{A^{-1}A}_{I} z = A^{-1}b$$

This simplifies to $Iz = A^{-1}b$. But $Iz = z$, and so $z = A^{-1}b$ (the one solution we already have).

Part (b) deals with the special case $b = 0$. To prove (b), observe that $A0 = 0$, so $x = 0$ is a solution to the equation $Ax = 0$. Part (a) shows this solution is unique. □

2.10.5. Chains of invertible matrices

Suppose we have two invertible $n \times n$ matrices A and B with inverses A^{-1} and B^{-1}, respectively. Is AB invertible? Yes.

LEMMA 2.6. *If A and B are invertible matrices of the same size, with inverses A^{-1} and B^{-1}, then AB is invertible and*

$$(AB)^{-1} = B^{-1}A^{-1}$$

PROOF. We simply verify that $B^{-1}A^{-1}$ satisfies the requirements of an inverse of AB. We have

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = AIA^{-1} = AA^{-1} = I$$

also

$$(B^{-1}A^{-1})(AB) = B^{-1}(A^{-1}A)B = B^{-1}B = I$$

Therefore, $B^{-1}A^{-1}$ satisfies the definition of an inverse of AB. □

The above lemma, with induction, can prove the following: Let $A_1, A_2, \ldots A_k$ be a collection of invertible matrices of the same size, with inverses $A_1^{-1}, A_2^{-1}, \ldots, A_k^{-1}$. Then the product $A_1A_2\cdots A_k$ is invertible with inverse

$$(A_1A_2\cdots A_{k-1}A_k)^{-1} = A_k^{-1}A_{k-1}^{-1}\cdots A_2^{-1}A_1^{-1} \tag{2.18}$$

This formula for the inverse of a product chain of invertible matrices is reminiscent of the formula for the transpose of a product chain. The next subsection establishes a deep connection between inverses and transposes.

2.10.6. The inverse of $A^\top$

LEMMA 2.7. *Let A be a square matrix with entries in some field $\mathbb{F}$. Then*

$$(A \text{ is invertible}) \iff \left(A^\top \text{ is invertible}\right)$$

Further, if A is invertible then

$$(A^\top)^{-1} = (A^{-1})^\top$$

PROOF. First suppose A is invertible. Then A^{-1} exists and

$$AA^{-1} = A^{-1}A = I$$

Taking transposes of the above gives

$$(A^{-1})^\top A^\top = A^\top (A^{-1})^\top = I^\top = I$$

where we have used the fact that I is symmetric, so $I^\top = I$. It follows that $(A^{-1})^\top$ satisfies the definition for an inverse of $A^\top$. □

LEMMA 2.8. *Let A be a $n \times n$ matrix with entries in some field $\mathbb{F}$.*

a) If A has a column of all zeros, then A is not invertible.

b) If A has a row of all zeros, then A is not invertible.

PROOF. To prove (a), suppose A has all zeros in some column $i \in \{1, \ldots, n\}$. Let e_i be the ith column of the $n \times n$ identity matrix. Then $e_i \neq 0$ and

$$Ae_i = (i\text{th column of } A) = 0$$

Thus, the homogeneous equation $Ax = 0$ has a nontrivial solution (that is, it has a solution that is not the all-zero vector). By Lemma 2.5(b), A cannot be invertible.

To prove (b), note that if A has an all-zero row then $A^\top$ has an all-zero column. Part (a) implies $A^\top$ is not invertible, so Lemma 2.7 implies A is not invertible. □

2.10.7. The inverse of a 2×2 matrix

It is not trivial to determine whether or not a given square matrix is invertible, particularly when the matrix is large. If the matrix *is* invertible, we want to compute its inverse. In general we need some computational procedure (the most common procedure is called *Gaussian elimination* and is described in Chapter 4). However, the special case of 2×2 matrices is easy. We can quickly determine whether or not a 2×2 matrix is invertible. If it is invertible, there is a simple closed form formula for the inverse.

LEMMA 2.9. *Let $\mathbb{F}$ be a field. Consider the following 2×2 matrix A*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \tag{2.19}$$

where $a, b, c, d \in \mathbb{F}$. Then A is invertible if and only if $ad - bc \neq 0$. If $ad - bc \neq 0$ then

$$A^{-1} = (ad - bc)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \tag{2.20}$$

Before proving the lemma, the expression on the right-hand-side of (2.20) needs some discussion: Since a, b, c, d are scalars in some given field $\mathbb{F}$, the expression $ad - bc$ is also a scalar in $\mathbb{F}$. Since we have assumed $ad - bc \neq 0$, the multiplicative inverse axiom ensures $(ad - bc)^{-1}$ exists (it is some other scalar in $\mathbb{F}$). Therefore, the expression on the right-hand-side of (2.20) multiplies the scalar $(ad - bc)^{-1}$ by a 2×2 matrix, a type of multiplication that is defined in Section 2.3 (it means that we multiply all 4 entries of the 2×2 matrix by that scalar). The 2×2 matrix itself uses scalars $-b$ and $-c$, which are known to exist in $\mathbb{F}$ by the additive inverse axiom (we already know $b, c \in \mathbb{F}$). For the 2×2 matrix A in (2.19), we define the *determinant of A*, written $\det(A)$, by

$$\det(A) = ad - bc$$

Lemma 2.9 shows that if A is a matrix of size 2×2 then

$$(A \text{ is invertible}) \iff \det(A) \neq 0$$

Exercises 2.63 and 2.65 explore properties of determinants of 2×2 matrices. Determinants for general $n \times n$ matrices are developed in Chapter 11.

PROOF. (Lemma 2.9) First suppose $ad - bc \neq 0$. We show the matrix in (2.20) satisfies the definition of an inverse of A. We have

$$\begin{aligned} \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{A} \left((ad - bc)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \right) &\overset{(a)}{=} (ad - bc)^{-1} \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \\ &= (ad - bc)^{-1} \begin{bmatrix} ad - bc & 0 \\ 0 & ad - bc \end{bmatrix} \\ &\overset{(b)}{=} I \end{aligned}$$

where equalities (a) and (b) pull out the scalar from the matrix multiplication. Similarly,

$$\left((ad-bc)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \right) \underbrace{\begin{bmatrix} a & b \\ c & d \end{bmatrix}}_{A} = I$$

Thus, we have satisfied the definition of an inverse of A.

For the reverse direction, suppose $ad - bc = 0$. We must show the matrix A is *not* invertible. This part is given in Exercise 2.22. □

Examples:

- For $\mathbb{F} = \mathbb{R}$ we have

$$\begin{bmatrix} 4 & -1 \\ 0 & 2 \end{bmatrix}^{-1} = 8^{-1} \begin{bmatrix} 2 & 1 \\ 0 & 4 \end{bmatrix} = \begin{bmatrix} 1/4 & 1/8 \\ 0 & 1/2 \end{bmatrix}$$

- For $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic we have

$$\begin{bmatrix} 1 & 1 \\ 1 & 0 \end{bmatrix}^{-1} = 1^{-1} \begin{bmatrix} 0 & -1 \\ -1 & 1 \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ 1 & 1 \end{bmatrix}$$

  which uses the fact that $-1 = 1$ and $1^{-1} = 1$. [Final answer uses only scalars $\{0, 1\}$.]

- For $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ with mod 5 arithmetic we have

$$\begin{bmatrix} 3 & 1 \\ 2 & 3 \end{bmatrix}^{-1} = 7^{-1} \begin{bmatrix} 3 & -1 \\ -2 & 3 \end{bmatrix} = 3 \begin{bmatrix} 3 & 4 \\ 3 & 3 \end{bmatrix} = \begin{bmatrix} 4 & 2 \\ 4 & 4 \end{bmatrix}$$

  which uses $-1 = 4$, $-2 = 3$, $7^{-1} = 2^{-1} = 3$. [Final answer uses only scalars $\{0, 1, 2, 3, 4\}$.]

2.11. Permutation matrices

A *permutation matrix* P is a square matrix with entries that are all either 0 or 1, and that has exactly one "1" entry on each row and column. Since all fields $\mathbb{F}$ contain scalars 0 and 1, permutation matrices make sense in every field. Every permutation matrix of size $n \times n$ can be formed by starting with the $n \times n$ identity matrix I and then permuting its rows. Equivalently, a permutation matrix can be formed by permuting the columns of I.

2.11.1. Counting permutation matrices

There are exactly $n!$ permutation matrices of size $n \times n$, one for each permutation of the integers in the set $\{1, \ldots, n\}$. This is because to create a $n \times n$ permutation matrix we need to make the following sequence of choices:

- Row 1: Choose which of the n columns to place the 1 (n choices).
- Row 2: Choose which of the remaining $(n-1)$ columns to place the 1 ($n-1$ choices).
- Row 3: Choose which of the remaining $(n-2)$ columns to place the 1 ($n-2$ choices).
- $\vdots$
- Row n: Place the 1 in the only remaining column (1 choice)

So the total number of choices is:

$$n(n-1)(n-2)\cdots(2)(1) = n!$$

2.11.2. Multiplying PA and AP

Let P be a $n \times n$ permutation matrix. Consider matrices

$$A \sim m \times n \quad , \quad B \sim n \times k$$

Then

- AP is a $m \times n$ matrix that permutes the n columns of A. Since $IP = P$, the way that P permutes the n columns of A is the same as how P permutes the n columns of I. For example, if the first column of P is e_j (the jth column of I) then the first column of AP is also the jth column of A.
- PB is a $n \times k$ matrix that permutes the n rows of B. Since $PI = P$, the way that P permutes the n rows of I is the same as how P permutes the n rows of B. For example, if the first row of P is $e_i^\top$ (the ith row of I) then the first row of PB is also the ith row of B.

1. A 2×2 example:

$$\underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{P}\begin{bmatrix} a & b \\ c & d \end{bmatrix} = \begin{bmatrix} c & d \\ a & b \end{bmatrix}$$

$$\begin{bmatrix} a & b \\ c & d \end{bmatrix}\underbrace{\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}}_{P} = \begin{bmatrix} b & a \\ d & c \end{bmatrix}$$

2. A 3×3 example:

$$\underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{P}\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} g & h & i \\ a & b & c \\ d & e & f \end{bmatrix}$$

$$\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{P} = \begin{bmatrix} b & c & a \\ e & f & d \\ h & i & g \end{bmatrix}$$

2.11.3. Permutation matrices are invertible

Permutation matrices are always invertible matrices because there is always another permutation matrix that can "undo" the permutations. It can be shown that the inverse of a permutation matrix P is equal to the transpose of P (see Exercises 2.37 and 2.38):

$$(PP^\top = P^\top P = I) \implies P^{-1} = P^\top$$

2.11.4. Row swap permutations

The most basic type of permutation matrix is one formed by swapping two distinct rows i and j of the identity matrix I. These permutation matrices are important in the theory of *Gaussian elimination* (discussed in Chapter 4). Left-multiplying a matrix A by this type of permutation matrix swaps rows i and j of A:

[swap rows 1 and 3]: $$\underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{P}\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \\ j & k & l \end{bmatrix} = \begin{bmatrix} g & h & i \\ d & e & f \\ a & b & c \\ j & k & l \end{bmatrix}$$

It follows that left-multiplying such a permutation matrix P by itself performs a row swap on itself to turn it back into the identity matrix:

$$(PP = P^2 = I) \implies P^{-1} = P$$

Thus, such a permutation matrix is its own inverse matrix. Also, all such row-swap matrices are symmetric, so $P^\top = P$, which is another reason why $P^{-1} = P$ for these matrices. It can be shown that every permutation matrix can be written as a product of row-swap permutation matrices.

2.12. Exercises

EXERCISE 2.1. *Give a specific example of matrices A, B, C (not all square, and none of them being an all-zero matrix), with appropriate size so that ABC makes sense. Compute AB, BC, $(AB)C$, and $A(BC)$.*

EXERCISE 2.2. *Let $A_1, \ldots, A_n$ be $m \times n$ matrices with entries in some field $\mathbb{F}$. Show that*

$$\left(\sum_{s=1}^{n} A_s\right)^\top = \sum_{s=1}^{n} A_s^\top$$

Specifically, fix i, j and show the (i, j) entry of the left-hand-side is the same as the (i, j) entry of the right-hand-side.

EXERCISE 2.3. *(Matrix Multiplication with fields $\mathbb{R}$ and $\mathbb{F}_3$)*

a) Compute the following three matrix products using the field $\mathbb{F} = \mathbb{R}$

$$\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\begin{bmatrix} 1 & 1 & 1 \\ 1 & 0 & 2 \end{bmatrix}, \begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 2 & 1 \end{bmatrix}$$

b) Repeat part (a), but assuming we use the field $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic.

EXERCISE 2.4. *(Matrix Multiplication with fields $\mathbb{R}$, $\mathbb{F}_3$, and $\mathbb{F}_5$)*

a) Compute the following two matrix products using the field $\mathbb{F} = \mathbb{R}$.

$$\begin{bmatrix} 0 & 1 & 1 \\ 0 & 2 & 0 \\ 2 & 2 & 2 \end{bmatrix}\begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 & 2 & 2 \end{bmatrix}\begin{bmatrix} 0 & 1 & 1 \\ 0 & 2 & 0 \\ 2 & 2 & 2 \end{bmatrix}$$

b) Repeat part (a) using field $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic.

c) Repeat part (a) using field $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ with mod 5 arithmetic.

EXERCISE 2.5. *(Associativity of matrix multiplication) Compute the following matrix product ABC over $\mathbb{F}_3 = \{0, 1, 2\}$ two ways: First, compute it using the order $A(BC)$. Second, compute it using the order $(AB)C$. Which is easier?*

$$\begin{bmatrix} 1 & 2 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & 1 & 1 & 2 & 0 & 1 \end{bmatrix}$$

EXERCISE 2.6. *(Solving) For this problem, you can either search over all possible solution vectors with a computer, or write out the system of three equations and solve by hand via algebraic manipulation. There is no need to do Gaussian elimination here. However, you can use Gaussian elimination if you know it.*

a) Find all solutions over the mod-2 field $\mathbb{F}_2 = \{0, 1\}$ *(there are* $2^4 = 16$ *vectors in* $\mathbb{F}_2^4$*):*

$$\begin{bmatrix} 1 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} w \\ x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$$

b) For the same matrix in part (a), find all solutions over the mod-3 field $\mathbb{F}_3 = \{0, 1, 2\}$ *(there are* $3^4 = 81$ *vectors in* $\mathbb{F}_3^4$*).*

EXERCISE 2.7. *Consider field* $\mathbb{F} = \mathbb{R}$*. Let* v_1, v_2, v_3, v_4 *be vectors in* $\mathbb{R}^m$*. Define the* $m \times 4$ *matrix* $V = [v_1\ v_2\ v_3\ v_4]$*.*

a) Design a 4×2 *matrix* A *such that the first column of* VA *is* $v_1 + v_2 - 8v_4$ *and the second column is* $v_2 + v_3$*.*

b) Design a 4×3 *matrix* B *so that the first column of* VB *is* v_3*, the second column is* $v_1 + v_2 + v_3$*, the third is* $-v_2 + 5v_4$*.*

c) Design a 4×2 *matrix* C *so that the first column of* VC *is* 0 *and the second column is* $v_1 - 27v_4$*.*

d) What are the sizes of matrices E *and* F *such that* $EVF \in \mathbb{R}$*?*

EXERCISE 2.8. *Consider* $\mathbb{F} = \mathbb{R}$*. Let* y_1, y_2, y_3, y_4 *be vectors in* $\mathbb{R}^m$*. Define the* $4 \times m$ *matrix* D *so that the* i*th row of* D *is* $y_i^\top$*, for* $i \in \{1, 2, 3, 4\}$*.*

a) Design a 2×4 *matrix* A *so that* AD *has first row* $8y_2^\top - 3y_3^\top$ *and second row* $y_1^\top + y_3^\top + 10y_4^\top$*.*

b) Design a 3×4 *matrix* B *so that* BD *has first row* $90y_2^\top$*, second row* $y_1^\top + y_4^\top$*, third row* $-7y_3^\top + 8y_4^\top$*.*

c) Design a 2×4 *matrix* C *so that* CD *has first row* $0^\top$ *and second row* $y_1^\top + y_2^\top - y_3^\top + y_4^\top$*.*

EXERCISE 2.9. *Using the mod-2 field* $\mathbb{F}_2 = \{0, 1\}$*, find a* 2×2 *matrix* B *such that* $AB = C$*, where* A *and* C *are given below:*

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

EXERCISE 2.10. *Using $\mathbb{F}_3 = \{0, 1, 2\}$ and mod 3 arithmetic, consider:*

$$A = \begin{bmatrix} 0 & 0 & 2 & 2 \\ 1 & 2 & 0 & 1 \\ 0 & 1 & 2 & 1 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 2 & 1 & 0 \\ 2 & 0 & 0 \end{bmatrix} \quad , \quad d = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$$

a) Compute AB and $B^\top A^\top$ (using mod 3 arithmetic).

b) Compute $d^\top d$ and $d^\top A$ (using mod 3 arithmetic).

c) Find two distinct solutions to $Ax = d$ (using mod 3 arithmetic). You can either play around with the equations, or you can write a computer program that searches over all 3^4 possible values of $x \in \{0, 1, 2\}^4$. Note that the matlab command `mod(A * x, 3)` *outputs the mod 3 value of the vector Ax.*

EXERCISE 2.11. *Compute AB, BC, $(AB)C$, and $A(BC)$ for the case $\mathbb{F} = \mathbb{R}$ and*

$$A = \begin{bmatrix} 1 \\ 0 \\ -1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}, \quad C = \begin{bmatrix} 4 \\ 5 \\ 6 \end{bmatrix}$$

EXERCISE 2.12. *(Matrix multiplication is distributive) Let A, B be matrices of the same size. Let C be a matrix with a number of rows equal to the number of columns of A.*

a) Show that

$$(A + B)C = AC + BC$$

b) Use transposes to show that if D, E are matrices of the same size, and if F is a matrix with number of columns equal to the number of rows of D, then

$$F(D + E) = FD + FE$$

EXERCISE 2.13. *(Multiplying matrices in binary field) The following matrices have entries in the binary field $\mathbb{F}_2 = \{0, 1\}$:*

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad G = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

a) Compute GA and Gb (using the binary field $\mathbb{F}_2$).

b) Find the set S_1 of all binary-valued column vectors $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ that solve the following (with field $\mathbb{F}_2$):

$$A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = b$$

c) Find the set S_2 of all binary-valued column vectors $x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ that solve the following (with field $\mathbb{F}_2$):

$$GA \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = Gb$$

d) Do we have $S_1 = S_2$? Why or why not?

EXERCISE 2.14. *(The equation $Ax = b$) Let $\mathbb{F}$ be a field. Suppose A is a $m \times n$ matrix and b is a column vector of size m, both having entries in $\mathbb{F}$. We want to find vectors $x \in \mathbb{F}^n$ that solve the equation $Ax = b$.*

a) Suppose y and z are vectors in $\mathbb{F}^n$ that solve $Ax = b$. Show that for all $t \in \mathbb{F}$, the vector $ty + (1-t)z$ also solves $Ax = b$.

b) Show that if $\mathbb{F} = \mathbb{R}$ then the number of solutions to $Ax = b$ is either 0, 1, or ∞.

EXERCISE 2.15. *(Matrix multiplication with reals) Using $\mathbb{F} = \mathbb{R}$ define*

$$A = \begin{bmatrix} 1 \\ 2 \\ -3 \\ 0 \\ 8 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 1 & 1 & 0 & 1 \end{bmatrix} \quad , \quad C = \begin{bmatrix} 4 & 2 \\ 0 & 1 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}$$

a) Compute AB and $(AB)C$.

b) Compute BC and $A(BC)$. Discuss.

EXERCISE 2.16. *(trace(AB) = trace(BA)) Let m, n be positive integers. Let $A = (a_{ij}) \sim m \times n$ and $B = (b_{ij}) \sim n \times m$. We want to use a sum-switching argument to show the multiplications AB and BA make sense and lead to square matrices that satisfy trace(AB) = trace(BA). All entries of the matrices are assumed to be in some common field.*

a) Show that AB and BA are both well defined and square. What are their sizes?

b) Define $(AB)_{ij}$ as the (i, j) entry of the matrix AB. Show that trace(AB) = trace(BA) by

filling in the (two or three) gaps of the following chain of equalities:

$$trace(AB) = \sum_{i=1}^{m} (AB)_{ii} = \ldots = \ldots = \sum_{j=1}^{n} (BA)_{jj} = trace(BA)$$

Make sure to write $(AB)_{ii}$ in terms of the a_{ij} and b_{ji} values using the definition of matrix multiplication.

c) Give a specific example.

EXERCISE 2.17. *(Complexity of computing $A^k x$) Let $A \sim n \times n$ with entries in some field $\mathbb{F}$. Fix $x \in \mathbb{F}^n$. Fix k as a positive integer. We want to compute $A^k x$. Note that multiplying two $n \times n$ matrices has complexity n^3 (in units of multiply-adds). Multiplying a $n \times n$ matrix and a $n \times 1$ vector has complexity n^2. A naive implementation would multiply A by itself $k-1$ times and then multiply x, which has complexity $(k-1)n^3 + n^2$ (too high). This problem considers two approaches, both of which have significantly lower complexity than $(k-1)n^3 + n^2$ (see also computer implementation in Exercise 2.52).*

- *Approach 1: Given A and x, compute:*

$$\begin{aligned} h_1 &= Ax \\ h_2 &= Ah_1 \\ h_3 &= Ah_2 \\ &\vdots \\ h_k &= Ah_{k-1} \end{aligned}$$

  and h_k is the desired result $A^k x$.

- *Approach 2: Assume $k = 2^r$ for some positive integer r. Given A and x, compute:*

$$\begin{aligned} A^2 &= (A)(A) \\ A^4 &= (A^2)(A^2) \\ A^8 &= (A^4)(A^4) \\ &\vdots \\ A^k &= (A^{k/2})(A^{k/2}) \end{aligned}$$

  and then use n^2 additional complexity to compute $(A^k)x$.

Intuitively, Approach 1 is better when "n is large relative to k" (since Approach 1 avoids the n^3 complexity of multiplying a matrix by a matrix), while Approach 2 is better when "k is large relative to n" (since Approach 2 has much fewer steps).

a) Find the total complexity for each of these two approaches. For which values of k, n is Approach 2 better than Approach 1? If $k = 1024$, for which values of n is Approach 2 better?

b) Suppose $k = 1000$ (not a power of 2). Modify Approach 2 and find the total complexity for computing $A^{1000}x$. Can you find values of n for which the modified approach is better than Approach 1? Hint: $1000 = 2^9 + 2^8 + 2^7 + 2^6 + 2^5 + 2^3$.

EXERCISE 2.18. *(Multiplying 2×2 and 3×3 upper triangular matrices) Use $\mathbb{F} = \mathbb{R}$. Perform the following multiplications of upper triangular matrices to show the result is again upper triangular.*

a)

$$\begin{bmatrix} 3 & -1 \\ 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 7 \\ 0 & 9 \end{bmatrix} =$$

b)

$$\begin{bmatrix} 5 & -1 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 2 \end{bmatrix}\begin{bmatrix} 1 & 7 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 2 \end{bmatrix} =$$

EXERCISE 2.19. *Let A be a square invertible matrix with entries in some field. Show that A^{-1} is also invertible and $(A^{-1})^{-1} = A$.*

EXERCISE 2.20. *Let A be a general 2×2 upper triangular matrix (with entries in some field).*

$$A = \begin{bmatrix} a & b \\ 0 & d \end{bmatrix}$$

a) Show that A is invertible if and only if $a \neq 0$ and $d \neq 0$.

b) Compute A^{-1} and show the result is again upper triangular.

c) Compute $(A^\top)^{-1}$. Observe that both $A^\top$ and $(A^\top)^{-1}$ are lower triangular.

EXERCISE 2.21. *(Multiplying upper triangular matrices) Let $\mathbb{F}$ be a field. Let A and B be $n \times n$ matrices with entries in $\mathbb{F}$. Suppose both A and B are upper triangular. We want to show that AB is also upper triangular. To do this, fix $i, j \in \{1, \ldots, n\}$ such that $i > j$. It suffices to show that the (i, j) entry of AB is zero.*

a) Show that the (i, j) entry of AB is zero. [Hint: Use the summation formula for the (i, j) entry and split the sum via $\sum_{w=1}^{n} = \sum_{w<i} + \sum_{w \geq i}$.]

b) Part (a) proves the desired result for upper triangular matrices. Use transposes to show the corresponding result for lower triangular matrices.

c) Using field $\mathbb{R}$, give a specific example of two upper triangular 2×2 matrices that multiply to an upper triangular 2×2 matrix.

d) Using field $\mathbb{R}$, give a specific example of two upper triangular 3×3 matrices that multiply to an upper triangular 3×3 matrix.

e) Using field $\mathbb{R}$, is the product of a $n \times n$ upper triangular matrix and a $n \times n$ lower triangular

matrix a diagonal matrix? Prove true, or give a counter-example.

EXERCISE 2.22. *(Invertibility of a* 2×2 *matrix) Let* $\mathbb{F}$ *be any field. Fix* $a, b, c, d \in \mathbb{F}$ *and define the* 2×2 *matrix*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Lemma 2.9 proved that A *is invertible whenever* $ad - bc \neq 0$*. Here we show the converse: If* $ad - bc = 0$*, then* A *is not invertible.*

a) Suppose $a = 0$*. Claim that* A *has either a column or row of all-zeros, so it cannot be invertible (by Lemma 2.8).*

b) Suppose $a \neq 0$*. Prove that the homogeneous equation* $Ax = 0$ *has a nontrivial solution (meaning a solution that is not the all-zero vector), so* A *is not invertible (by Lemma 2.5(b)).*

EXERCISE 2.23. *Solving* 2×2 *systems. [If there is no solution, explain why.]*

a) Solve using $\mathbb{F} = \mathbb{R}$*:*

$$\begin{bmatrix} 3 & 4 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

b) Solving using $\mathbb{F} = \mathbb{R}$*:*

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

c) Solving using $\mathbb{F}_2 = \{0, 1\}$ *and mod 2 arithmetic:*

$$\begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

d) Solving using $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ *and mod 5 arithmetic:*

$$\begin{bmatrix} 1 & 3 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

EXERCISE 2.24. *(Finding* A^{-1} *for a* 2×2*) Define*

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

a) Find A^{-1} *and* $(A^\top)^{-1}$ *using* $\mathbb{F} = \mathbb{R}$*.*

b) Find A^{-1} *and* $(A^\top)^{-1}$ *using* $\mathbb{F} = \mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ *with mod 5 arithmetic.*

EXERCISE 2.25. *(Again finding A^{-1} for a 2×2) Define*

$$A = \begin{bmatrix} 3 & 3 \\ 2 & 1 \end{bmatrix}$$

a) Find A^{-1} and $(A^\top)^{-1}$ using $\mathbb{F} = \mathbb{R}$.

b) Find A^{-1} and $(A^\top)^{-1}$ using $\mathbb{F} = \mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ with mod 5 arithmetic.

EXERCISE 2.26. *(Another finding A^{-1} for a 2×2) Define*

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

a) Find A^{-1} and $(A^\top)^{-1}$ using $\mathbb{F} = \mathbb{R}$.

b) Find A^{-1} and $(A^\top)^{-1}$ using $\mathbb{F} = \mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic.

EXERCISE 2.27. *($A^\top A$ is symmetric and positive semidefinite) Suppose A is a $m \times n$ matrix with entries in some field $\mathbb{F}$.*

a) Show that $A^\top A$ is a symmetric $n \times n$ matrix.

b) Fix $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$. Treating y as a column vector, show that $y^\top y$ is a real number that is nonnegative.

c) Let A be $m \times n$ with real entries. We already know from part (a) that $A^\top A$ is a symmetric $n \times n$ matrix. We want to show $A^\top A$ is symmetric and positive semidefinite*, meaning that it is symmetric and*

$$x^\top A^\top A x \geq 0 \quad \forall x \in \mathbb{R}^n$$

Start your proof by writing "Fix $x \in \mathbb{R}^n$. We want to show $x^\top A^\top A x \geq 0$." Continue your proof by writing the expression $x^\top A^\top A x$ in a form that is compatible with part (b). You may want to define a vector y in terms of things that have already been instantiated. [Note: This basic exercise is a key to the theory of singular values *of a real $m \times n$ matrix, discussed in Chapter 16.]*

EXERCISE 2.28. *(Computer experiment with trace) Let $A \sim 3 \times 2$ and $B \sim 2 \times 3$.*

a) Compute the sizes of AB, BA, $B^\top A^\top$, $A^\top B^\top$.

b) Use MATLAB *(or some other computer software) to generate a random 3×2 matrix A and a random 2×3 matrix B with entries that are independent and uniformly distributed in $[0, 1]$. Compute $\mathrm{trace}(AB)$, $\mathrm{trace}(BA)$, $\mathrm{trace}(B^\top A^\top)$, $\mathrm{trace}(A^\top B^\top)$. (You are free to cut and paste the* MATLAB *commands below.)*

c) Use the theory of Chapter 2.7 to explain your findings in part (b).

Listing 2.1. MATLAB defining A and B for Exercise 2.28

```
%This program computes random matrices A, B
```

```
A=rand(3,2);
B=rand(2,3);
trace(A*B)
trace(B*A)
trace(B'*A')
trace(A'*B')
```

EXERCISE 2.29. *(2 × 2 and 3 × 3 permutation matrix)*

a) Find all 2 × 2 *permutation matrices. How many are there? What are their inverse matrices?*

b) Find all 3 × 3 *permutation matrices. How many are there?*

c) Find the inverse of the following 3 × 3 *permutation matrix:*

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

Verify that $PP^{-1} = P^{-1}P = I$.

EXERCISE 2.30. *(Multiplication by a permutation matrix) Define*

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 8 & 8 & 8 & 8 \\ 5 & 6 & 7 & 8 \\ 9 & 9 & 9 & 9 \end{bmatrix}, \quad P_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} \quad P_2 = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The matrices P_1, P_2 *are permutation matrices that correspond to row-swap matrices (they are formed by swapping two rows of I).*

a) Compute P_1A *and* P_1P_2.

b) Compute AP_1 *and* P_2P_1.

c) Compute P_2A *and* P_1P_2A.

d) Compute P_1^{-1}, P_2^{-1}, *and* $(P_1P_2)^{-1}$. *Verify that in these cases, the inverse of a permutation matrix is the transpose.*

EXERCISE 2.31. *(Permutation as a sequence of row swaps) Consider the* 3 × 3 *permutation matrix:*

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

a) Write P as a product of row-swap permutation matrices P_i *(where each* P_i *is formed by swapping two rows of the* 3 × 3 *identity matrix I).*

b) Find P^{-1}.

EXERCISE 2.32. *(5×5 permutation as a sequence of row swaps) Consider the 5×5 permutation matrix:*

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

a) Write P as a product of row-swap permutation matrices P_i (where each P_i is formed by swapping two rows of the 5×5 identity matrix I).

b) Find P^{-1}.

EXERCISE 2.33. *Let $\mathbb{F} = \mathbb{R}$. Define*

$$A = \begin{bmatrix} 4 & 5 \\ 2 & 3 \end{bmatrix}$$

a) Find a matrix B such that $AB = I$.

b) Find a matrix C such that $CA = I$.

EXERCISE 2.34. *Suppose A is a square matrix with entries in some field $\mathbb{F}$. Suppose B and C are square matrices of the same size as A (and with entries in $\mathbb{F}$). Suppose $AB = I$ and $CA = I$. Show that $B = C$ and A is invertible with $A^{-1} = B$.*

EXERCISE 2.35. *Suppose A and B are square matrices of the same size.*

a) Suppose $BA = I$. Show that $Ax = 0$ has only the trivial solution $x = 0$.

b) Suppose $Ax = b$ has a solution for all column vectors b of appropriate size. Use this to build a matrix C such that $AC = I$. Hint: Build the columns of C one at a time.

EXERCISE 2.36. *(Distributing a scalar across a matrix) Let $\mathbb{F}$ be a field. Fix $c, d \in \mathbb{F}$. Let A and B be $m \times n$ matrices with entries in $\mathbb{F}$.*

a) Show that $c(A + B) = cA + cB$.

b) Show that $(c + d)A = cA + dA$.

EXERCISE 2.37. *(Inverting a permutation matrix) Let $P = (p_{ij})$ be a $n \times n$ permutation matrix. We want to show that $P^{-1} = P^{\top}$.*

a) Show that for all $i \in \{1, \ldots, n\}$ we have $\sum_{u=1}^{n} p_{iu} = 1$.

b) Prove that for $i, j, u \in \{1, \ldots, n\}$ we have

$$p_{iu}p_{ju} = \begin{cases} 0 & \text{if } i \neq j \\ p_{iu} & \text{if } i = j \end{cases}$$

c) Fix (i, j) with $i, j \in \{1, \ldots, n\}$. Show that $(PP^\top)_{ij} = (I)_{ij}$ where I is the $n \times n$ identity matrix. Conclude $PP^\top = I$. By reversing the roles of P and $P^\top$, argue that we also have $P^\top P = I$.

EXERCISE 2.38. *(Alternative proof of inverting a permutation matrix) Fix $n \in \mathbb{N}$.*

a) Fix $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Let P_{swap} be the $n \times n$ row-swap permutation matrix that swaps rows i and j. Explain why $P_{swap}^{-1} = P_{swap}$. Explain why P_{swap} is symmetric (so $P_{swap}^\top = P_{swap} = P_{swap}^{-1}$.

b) Every $n \times n$ permutation matrix P can be written as the product of row-swap permutation matrices:

$$P = P_k P_{k-1} \cdots P_2 P_1$$

for some $k \in \mathbb{N}$, where P_i is a row-swap permutation matrix for each $i \in \{1, \ldots, k\}$. Use this with formulas (2.18) *and* (2.15) *to prove that $P^{-1} = P^\top$.*

EXERCISE 2.39. *(Distributing across many matrices) This exercise uses induction to prove the matrix distribution formula* (2.12)*. Let A be a matrix with n columns. In this problem, all matrices A and B_i have entries in some common field $\mathbb{F}$. The matrices B_i are all assumed to be the same size and to have n rows.*

a) Use formula (2.7) *to prove that $A(B_1 + B_2 + B_3) = AB_1 + AB_2 + AB_3$. [Hint: $B_1 + B_2 + B_3 = (B_1 + B_2) + B_3$.]*

b) Fix $k \in \mathbb{N}$. Assume the matrix distribution rule works for sums of size k, so $A\left(\sum_{i=1}^k B_i\right) = \sum_{i=1}^k AB_i$ always. Now prove the result for $k + 1$.

c) Discuss how the formula $(B + D)C = BC + DC$ can be similarly extended.

EXERCISE 2.40. *(Example of integer equation with no integer solution) Give an example of a matrix A with integer entries and a vector b with integer entries such that $Ax = b$ has at least one solution x (considering the real field), but no all-integer solution x exists. Show that your solution x has all rational components.*

EXERCISE 2.41. *(Solutions in $\mathbb{Z}$ and $\mathbb{F}_p$) Let A be a $m \times n$ matrix with integer entries. Let $b \in \mathbb{R}^m$ be a vector with integer entries.*

a) Using the real field, suppose $Ax = b$ has a solution x with all integer components. Let p be a prime and define $\overline{A}$, $\overline{x}$, $\overline{b}$ by taking the mod p values of all entries in A, x, b (so, in MATLAB *notation, $\overline{A} =$ mod (A, p)). Explain why, in the mod p field, we have*

$$\overline{A}\overline{x} = \overline{b} \pmod{p}$$

Discuss the fact that if components of A and b are already *in the set $\{0, 1, \ldots, p-1\}$, and if the equation $Ax = b$ has an integer solution x (using the field $\mathbb{R}$), then it also has a solution x using the field $\mathbb{F}_p$.*

b) Solve the equation below using field $\mathbb{R}$ to find $x \in \mathbb{R}^3$. You should find that x has all integer

components. Then solve in $\mathbb{F}_{11} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$ *with mod* 11 *arithmetic:*

$$\begin{bmatrix} 1 & 2 & 2 \\ 2 & 4 & 0 \\ 0 & 1 & 10 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 6 \\ 0 \\ 10 \end{bmatrix}$$

EXERCISE 2.42. *(Solving a system of three equations and three unknowns) Let* $\mathbb{F}$ *be a general field. Define* $2 = 1 + 1$. *Define* $3 = 1 + 1 + 1$. *Consider*

$$\underbrace{\begin{bmatrix} 1 & 1 & 1 \\ 2 & 0 & 1 \\ 2 & 1 & 3 \end{bmatrix}}_{A} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

a) Assume $3 \neq 0$. *Solve the above equation to find* (x, y, z). *You should find exactly one solution, and your answer should be in terms of* 3^{-1}.

b) Give the solution for the case $\mathbb{F} = \mathbb{R}$. *Then, give the solution for the case* $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ *with mod 5 arithmetic.*

c) Consider the same equation under field $\mathbb{F}_3 = \{0, 1, 2\}$ *with mod 3 arithmetic. In this case, we have* $3 = 0$ *and the final entry of the matrix A changes from* 3 *to* 0. *Is there a solution?*

EXERCISE 2.43. *(Basic transposes) Consider the binary field* $\mathbb{F}_2 = \{0, 1\}$ *with mod 2 arithmetic. Matrices A, B, C have entries in* $\mathbb{F}_2$ *and*

- $B^\top A^\top = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$
- $C = \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 0 \end{bmatrix}$

a) Compute ABC.

b) Give two different possibilities for A, B.

c) What are the possible sizes of matrices A, B?

EXERCISE 2.44. *(Invertible and not invertible with mod 2) For this exercise, use* $\mathbb{F}_2 = \{0, 1\}$ *with mod 2 arithmetic.*

a) Give two different 3×3 *matrices that are not invertible.*

b) Give two different 3×3 *matrices that are invertible.*

c) Give two different 3×2 *matrices with linearly independent columns.*

d) Give two different 3×2 *matrices with linearly dependent columns.*

EXERCISE 2.45. *(Block matrices) Let A be a 4×4 invertible matrix with entries in some field $\mathbb{F}$. Assume*

$$A = \begin{bmatrix} C \\ D \end{bmatrix} \quad , \quad A^{-1} = \begin{bmatrix} E & F \end{bmatrix}$$

where $C, D \sim 2 \times 4$ and $E, F \sim 4 \times 2$.

a) Compute CE, CF, DE, DF.

b) Give a computer-generated example with a random 4×4 matrix. For your example, you should compute CE, CF, DE, DF. You are free to copy and paste the MATLAB *below.*

c) For your same example from (b), compute EC, FC, ED, FD.

Listing 2.2. MATLAB for Exercise 2.45

```
A=rand(4,4)
Ainv = A^(-1)

C=A(1:2,:)
D=A(3:4,:)
E=Ainv(:,1:2)
F=Ainv(:,3:4)
```

EXERCISE 2.46. *(Block matrices with field $\mathbb{F}_{101}$) Let $A \sim 3 \times 3$ be an invertible matrix with entries in $\mathbb{F}_{101} = \{0, 1, \ldots, 100\}$:*

$$A = \begin{bmatrix} 5 & a & b \\ 3 & c & 0 \\ 1 & 1 & d \end{bmatrix} \quad , \quad A^{-1} = \begin{bmatrix} f & g & h \\ p & q & r \\ u & v & w \end{bmatrix}$$

where the variables a, b, c, d, f, g, and so on, represent unknown scalars. You should give answers in terms of these unknown scalars.

a) Using scalars $f, g, h, p, q, r, u, v, w$, find a matrix B such that (with mod 101 arithmetic):

$$B \begin{bmatrix} 5 & a \\ 3 & c \\ 1 & 1 \end{bmatrix} = I$$

(First find the required sizes of the square identity matrix I and the nonsquare matrix B.)

b) Using scalars $f, g, h, p, q, r, u, v, w$, find a matrix C such that (with mod 101 arithmetic):

$$C \begin{bmatrix} 5 & 1 \\ a & 1 \\ b & d \end{bmatrix} = I$$

EXERCISE 2.47. *(Building an inverse matrix with field $\mathbb{R}$) With the real field, define*

$$A = \begin{bmatrix} 3 & 1 & 2 \\ 0 & 3 & 9 \\ -1 & 1 & 0 \end{bmatrix}, e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

a) Find a vector $q \in \mathbb{R}^3$ such that $Aq = e_1$.

b) Find a vector $r \in \mathbb{R}^3$ such that $Ar = e_2$.

c) Find a vector $s \in \mathbb{R}^3$ such that $As = e_3$.

d) Find A^{-1} using your answers in parts (a)-(c). Verify that $AA^{-1} = A^{-1}A = I$.

EXERCISE 2.48. *(Building an inverse matrix with field $\mathbb{F}_3$) With field $\mathbb{F}_3 = \{0, 1, 2\}$ and mod 3 arithmetic define*

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 0 & 2 & 2 \\ 0 & 0 & 2 \end{bmatrix}, e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

a) Find a vector $q \in \mathbb{F}_3^3$ such that $Aq = e_1$.

b) Find a vector $r \in \mathbb{F}_3^3$ such that $Ar = e_2$.

c) Find a vector $s \in \mathbb{F}_3^3$ such that $As = e_3$.

d) Find A^{-1} using your answers in parts (a)-(c). Verify that $AA^{-1} = A^{-1}A = I$.

EXERCISE 2.49. *(Building an inverse matrix with field $\mathbb{F}_2$) With field $\mathbb{F}_2 = \{0, 1\}$ and mod 2 arithmetic define*

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

a) Find a vector $q \in \mathbb{F}_2^3$ such that $Aq = e_1$.

b) Find a vector $r \in \mathbb{F}_2^3$ such that $Ar = e_2$.

c) Find a vector $s \in \mathbb{F}_2^3$ such that $As = e_3$.

d) Find A^{-1} using your answers in parts (a)-(c). Verify that $AA^{-1} = A^{-1}A = I$.

EXERCISE 2.50. *(Building an inverse matrix with field $\mathbb{F}_3$) With field $\mathbb{F}_3 = \{0, 1, 2\}$ and mod 3 arithmetic define*

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}, e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

a) Find a vector $q \in \mathbb{F}_3^3$ such that $Aq = e_1$.
b) Find a vector $r \in \mathbb{F}_3^3$ such that $Ar = e_2$.
c) Find a vector $s \in \mathbb{F}_3^3$ such that $As = e_3$.
d) Find A^{-1} using your answers in parts (a)-(c). Verify that $AA^{-1} = A^{-1}A = I$.

EXERCISE 2.51. *(Commuting matrices) Let A be a $n \times n$ matrix with entries in some field. Let I be the $n \times n$ identity matrix (with that field). Let c, d be scalars.*
a) Show that $(A - cI)$ and $(A - dI)$ commute.
b) Give a specific example (with some particular integer $n \geq 2$).

EXERCISE 2.52. *(Markov chain computation) We implement the matrix multiplication approaches of Exercise 2.17 for the problem of computing probabilities in a Markov chain. By measuring elapsed computing time, Approaches 1 and 2 of Exercise 2.17 both perform significantly better than the naive approach described there. However, for* MATLAB*, elapsed times for Approaches 1 and 2 are better than what is predicted by the theoretical complexity in Exercise 2.17. This is due to implementation details such as parallel computing. [For example, the first two columns of a matrix multiplication AB can be computed simultaneously by two* parallel processors*, one given the task of computing Ab_1 and the other given the task Ab_2 (where b_1 and b_2 are the first two columns of B). See elapsed time for computing P^2 versus Px in Fig. 2.1.]*

Consider a discrete time Markov chain $X(t)$ with states $\{1, \ldots, n\}$ and transition matrix $P = (p_{ij})$, where $P[X(t+1) = j | X(t) = i] = p_{ij}$. Here, $p_{ij} \geq 0$ for all $i, j \in \{1, \ldots, n\}$ and $\sum_{j=1}^{n} p_{ij} = 1$ for all $i \in \{1, \ldots, n\}$. Let $\pi[t]$ denote the $n \times 1$ column vector with components $\pi_i[t] = P[X(t) = i]$ for $i \in \{1, \ldots, n\}$ and $t \in \{0, 1, 2, \ldots\}$. Assume the system starts in state 1 (with probability 1), so $\pi[0] = (1, 0, 0, \ldots, 0)$. The probabilities are

$$\pi[t]^\top = \pi[t-1]^\top P = \pi[0]^\top P^t \quad \forall t \in \{1, 2, 3, \ldots\}$$

It is well known that (under irreducible and aperiodic conditions)

$$\lim_{t \to \infty} \pi[t] = \pi$$

where π is the unique steady state probability vector that satisfies $\pi^\top = \pi^\top P$. Define $x = \pi[0]$ and

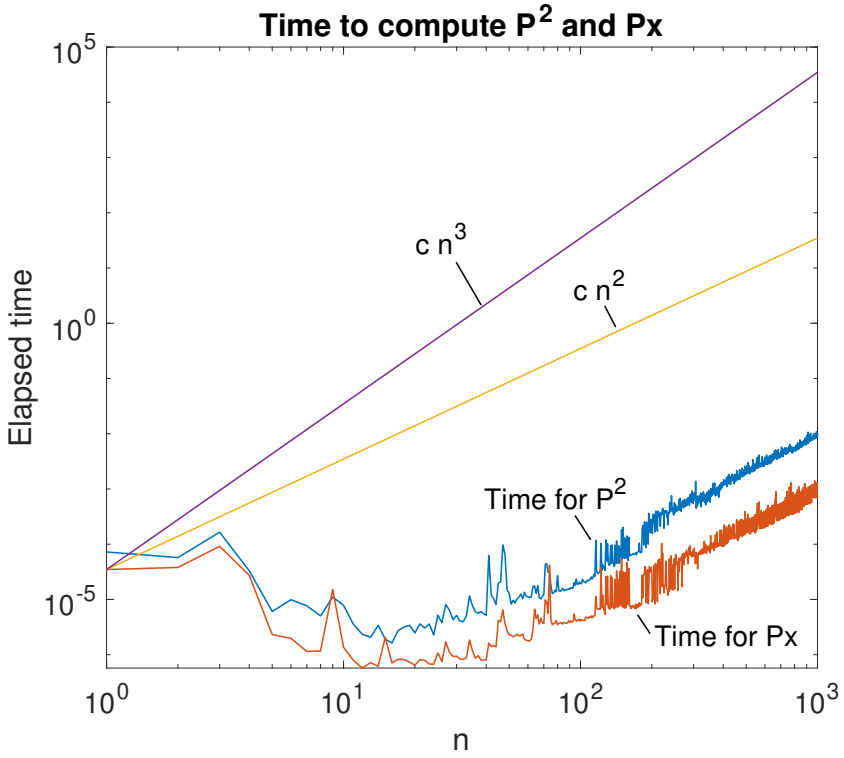


FIGURE 2.1. Loglog plot of empirical time for MATLAB to compute Px (complexity n^2) and P^2 (complexity n^3) for a random transition probability matrix P of size $n \times n$, and random probability mass vector x of size $n \times 1$. The theoretical complexity curves cn^2 and cn^3 are plotted with c chosen as a constant that matches the first data point for the empirical curve for elapsed time of Px. Hardware and software implementation details such as fast memory storage and parallel computing modes create a discrepancy between the measured empirical times and the scaled theoretical complexity cn^2 and cn^3. This relates to Exercises 2.17 and 2.52. The MATLAB program for producing this plot is below.

$A = P^\top$. *We want to compute* $A^k x$ *for large integer values of* k *(where* $(A^k x)^\top = \pi^\top[0]P^k = \pi[k]^\top$*). Assume* $k = 2^r$ *for some positive integer* r.

Write a computer program to test the elapsed time under the two approaches of Exercise 2.17, together with "Approach 0" which is the naive approach with high complexity $(k-1)n^3 + n^2$ *described there. Consider various* n, r*, and use a random* P *matrix. You are free to cut and paste the* MATLAB *program below. [Make sure not to copy the page number of this page!] You should include the cases* $(r = 10, n = 20)$*,* $(r = 10, n = 500)$*,* $(r = 10, n = 1000)$*. Also see if the output* $\pi[k]$ *is near steady-state by testing* $\pi[k]^\top \approx \pi[k]^\top P$*, where* $\pi[k] = A^k x$*. Discuss how the measured elapsed time does not account for the additional energy cost of parallel computing.*

Listing 2.3. MATLAB for Exercise 2.52.

```
%This program compares Approaches 0, 1, 2 for the Markov computation.
n=20
r=10;
k=2^r;
%Make random Markov transition prob matrix
B=rand(n,n);
P=zeros(n,n);
for i=1:n
  row = B(i,:);
  row = row./sum(row);  %Make all rows sum to 1
  P(i,:)=row;
end

x=zeros(n,1);
x(1)=1;
A=P';
%%%Approach 0
tic
B=A;
for i=2:k
  B=B*A;
end
h0=B*x;
ElapsedTime0=toc
%%%

%%%Approach 1
tic
h=A*x;
for i=2:k
  h=A*h;
end
```

```
h1 = h;
ElapsedTime1=toc
%%%

%%%Approach 2
tic
B=A;
for i=1:r
  B=B*B;
end
h2 = B*x;
ElapsedTime2=toc
%%%
```

Listing 2.4. MATLAB for producing Fig. 2.1.

```
nmax = 1000;
E1=[];
E2=[];
nvec=[];

for n=1:nmax
  nvec =[nvec;n];
  %Make random Markov transition prob matrix
  B=rand(n,n);
  P=zeros(n,n);
  for i=1:n
    row = B(i,:);
    row = row./sum(row);  %Make all rows sum to 1
    P(i,:)=row;
  end
  x=rand(n,1);
  x=x./sum(x);

  %%%Compute P^2%%%%%%
  tic
  P*P;
  elapse1=toc;
  E1 = [E1;elapse1];
  %%%%%%%%%%%%%%%%%%%%%%

  %%%Compute Px%%%%%%%%
  tic
  P*x;
```

```
  elapse2 = toc;
  E2 = [E2;elapse2];
  %%%%%%%%%%%%%%%%%%%%%%
end
c=E2(1);
loglog(nvec, E1, nvec, E2, nvec, c.*nvec.*nvec, nvec, c.*nvec.^3)
xlabel('n')
ylabel('Elapsed time');
```

EXERCISE 2.53. *(Determining an unknown function) Consider* $f : \mathbb{R}^2 \to \mathbb{R}^2$ *given by*

$$f(x_1, x_2) = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} x_1^2 \\ x_1 + x_2 \end{bmatrix} \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

where A *is an unknown* 2×2 *matrix with real entries. We sample the function at the vectors* $\{(2,1), (-1,3)\}$ *to obtain the following data:*

$$f(2,1) = \begin{bmatrix} 8 \\ 1 \end{bmatrix} \quad , \quad f(-1,3) = \begin{bmatrix} 5 \\ 5 \end{bmatrix}$$

Find A*.*

EXERCISE 2.54. *(Another unknown function) Consider* $f : \mathbb{R}^2 \to \mathbb{R}^2$ *given by*

$$f(x_1, x_2) = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} (x_1 + x_2)^2 \\ x_1 - 2x_2 \end{bmatrix} \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

where A *is an unknown* 2×2 *matrix with real entries.*

a) We sample the function at the vectors $\{(2,1), (1,2)\}$ *to obtain the following data:*

$$f(2,1) = \begin{bmatrix} -3 \\ 2 \end{bmatrix} \quad , \quad f(1,2) = \begin{bmatrix} 0 \\ 4 \end{bmatrix}$$

Find A*.*

b) We lose the data $f(2,1)$ *and* $f(1,2)$*. To compensate, we sample* f *at two new points* $\{(1,1), (5,5)\}$ *to obtain* $f(1,1)$ *and* $f(5,5)$*. Can we still find* A*? Explain.*

EXERCISE 2.55. *(Sampling an unknown function) Consider* $f : \mathbb{R}^2 \to \mathbb{R}^2$ *given by*

$$f(x_1, x_2) = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} e^{(x_1 + x_2)} \\ x_1 + 4x_2 \end{bmatrix} \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

where A is an unknown 2×2 matrix with real entries; $e^t = \exp(t)$ denotes the exponential function for $t \in \mathbb{R}$. Give two points $(x_1, x_2), (y_1, y_2) \in \mathbb{R}^2$ at which to sample f so that we can surely determine A from knowing $f(x_1, x_2)$, $f(y_1, y_2)$. Explain how we find A.

EXERCISE 2.56. *(A function with 5 unknown parameters) Consider $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by*

$$f(x_1, x_2) = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \theta \begin{bmatrix} x_1^2 - x_2 \\ x_1 + x_2 \end{bmatrix} \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

where A is an unknown 2×2 matrix with real entries, and $\theta \in \mathbb{R}$ is an unknown constant. Give three points $(x_1, x_2), (y_1, y_2), (z_1, z_2) \in \mathbb{R}^2$ at which to sample f so that we can surely determine A and θ from knowing $f(x_1, x_2)$, $f(y_1, y_2)$, $f(z_1, z_2)$. Explain how we can find A and θ.

EXERCISE 2.57. *(A function with 6 unknown parameters) Consider $f : \mathbb{R}^2 \to \mathbb{R}^2$ given by*

$$f(x_1, x_2) = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} \theta_1 & 0 \\ 0 & \theta_2 \end{bmatrix} \begin{bmatrix} x_1^2 - x_2^2 \\ 4x_2 \end{bmatrix} \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

where A is an unknown 2×2 matrix with real entries, and $\theta_1, \theta_2 \in \mathbb{R}$ are unknown constants. Give three points $(x_1, x_2), (y_1, y_2), (z_1, z_2) \in \mathbb{R}^2$ at which to sample f so that we can surely determine A, θ_1, θ_2 from knowing $f(x_1, x_2)$, $f(y_1, y_2)$, $f(z_1, z_2)$. Explain.

EXERCISE 2.58. *(Permutation matrices) Consider the real field with these matrices:*

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 4 & 5 \\ 8 & 8 & 8 & 8 & 8 \end{bmatrix}, \quad B = \begin{bmatrix} 8 & 8 & 8 & 8 & 8 \\ 1 & 2 & 2 & 4 & 5 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad C = \begin{bmatrix} 1 & 2 & 2 & 4 & 5 \\ 8 & 8 & 8 & 8 & 8 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

a) Find a 3×3 permutation matrix P_1 such that $P_1 A = B$.

b) Find a 3×3 permutation matrix P_2 such that $P_2 B = C$.

c) Define $P = P_2 P_1$. Compute P. Show that P is a permutation matrix and $PA = C$.

d) Compute $P_1^\top P_1$, $P_2^\top P_2$, $P^\top P$.

e) Compute $P_1^{-1}, P_2^{-1}, P^{-1}$.

EXERCISE 2.59. *(Symmetric matrices – True or false) Let A and B be $n \times n$ symmetric matrices. True or false: AB is symmetric. Prove true, or give a counter-example.*

EXERCISE 2.60. *($\det(A)$ and A^{-1} for 2×2 matrix) Consider*

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 4 \end{bmatrix}$$

a) Compute $\det(A)$ *assuming entries of A are in field* $\mathbb{R}$. *If* $\det(A) \neq 0$, *find* A^{-1}.

b) Compute $\det(A)$ *assuming entries of A are in field* $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ *with mod* 5 *arithmetic. If* $\det(A) \neq 0$, *find* A^{-1}.

c) Compute $\det(A)$ *assuming entries of A are in field* $\mathbb{F}_7 = \{0, 1, 2, 3, 4, 5, 6\}$ *with mod* 7 *arithmetic. If* $\det(A) \neq 0$, *find* A^{-1}.

EXERCISE 2.61. *(Invertible matrices in the binary field) Consider* 2×2 *matrices in the binary field* $\mathbb{F}_2 = \{0, 1\}$. *Such matrices either have* $\det(A) = 0$ *or* $\det(A) = 1$.

a) How many such matrices are there?

b) How many such matrices are not invertible (so $\det(A) = 0$*)? List them all.*

c) How many such matrices are invertible (so $\det(A) = 1$*)? List them all.*

d) Suppose we randomly choose such a matrix, equally likely over all possibilities in part (a). What is the probability our random matrix is invertible? Verify the answer matches the general formula in Chapter 5.9.2.

EXERCISE 2.62. *(Inverse of* 2×2*) Using field* $\mathbb{R}$, *find inverses for these matrices:*

$$A = \begin{bmatrix} -1 & 4 \\ 3/2 & 2 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 3 & 3 \\ 3 & 1 \end{bmatrix} \quad , \quad C = \begin{bmatrix} 9 & -1 \\ 5 & -2 \end{bmatrix}$$

EXERCISE 2.63. *(Determinants for* 2×2*) Consider these matrices with entries in a field* $\mathbb{F}$*:*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad , \quad B = \begin{bmatrix} c & d \\ a & b \end{bmatrix} \text{ [swaps rows of } A\text{]} \quad , \quad C = \begin{bmatrix} b & a \\ d & c \end{bmatrix} \text{ [swaps columns of } A\text{]}$$

a) Show that $\det(B) = -\det(A)$. *Argue that A is invertible if and only if B is invertible.*

b) Show that $\det(C) = -\det(A)$. *Argue that A is invertible if and only if C is invertible.*

c) Find $A^\top$ *and show that* $\det(A^\top) = \det(A)$.

d) Fix $\theta \in \mathbb{F}$. *Show that* $\det(\theta A) = \theta^2 \det(A)$.

EXERCISE 2.64. *(Conditions for invertibility) Consider the following matrices with entries in some field* $\mathbb{F}$*:*

$$A = \begin{bmatrix} 1 & a \\ 0 & 1 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 1 & b \\ 1 & 1 \end{bmatrix} \quad , \quad C = \begin{bmatrix} 1 & 1 \\ c & d \end{bmatrix}$$

a) Provide a necessary and sufficient condition on a for A to be invertible.

b) Provide a necessary and sufficient condition on b for B to be invertible.

c) Provide a necessary and sufficient condition on (c, d) *for C to be invertible.*

EXERCISE 2.65. *(*$\det(AB) = \det(A)\det(B)$*) Let* $\mathbb{F}$ *be a field. Define the following* 2×2 *matrices*

with entries in $\mathbb{F}$*:*

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \quad , \quad B = \begin{bmatrix} e & f \\ g & h \end{bmatrix}$$

a) Compute AB*.*

b) Show that $\det(AB) = \det(A)\det(B)$*.*

c) Show that $\det(A^3) = \det(A)^3$*.*

d) Fix $n \in \mathbb{N}$*. Show that* $\det(A^n) = \det(A)^n$*. [Use induction: Suppose for some* $k \in \mathbb{N}$ *that* $\det(A^k) = \det(A)^k$ *(we know it works for* $k = 2$*). Prove* $\det(A^{k+1}) = \det(A)^{k+1}$*.]*

e) Suppose A *is invertible. Show that* $\det(A^{-1}) = 1/\det(A)$*.*

EXERCISE 2.66. *(Invertible* 3×3*) Using field* $\mathbb{R}$*, consider*

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 4/3 \end{bmatrix} \quad , \quad C = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 4/3 \end{bmatrix} \quad , \quad D = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 2 & 0 \\ 0 & 0 & 4/3 \end{bmatrix}$$

a) Find a matrix B_1 *such that* $AB_1 = I$*. Show that we also have* $B_1A = I$*, so* $B_1 = A^{-1}$*.*

b) Find a matrix B_2 *such that* $CB_2 = I$*. Show that we also have* $B_2C = I$*, so* $B_2 = C^{-1}$*.*

c) Find a matrix B_3 *such that* $DB_3 = I$*. Show that we also have* $B_3D = I$*, so* $B_3 = D^{-1}$*.*

EXERCISE 2.67. *(Another invertible* 3×3*) Using field* $\mathbb{R}$*, consider*

$$A = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 9/5 & 0 \\ 0 & 0 & 5 \end{bmatrix} \quad , \quad C = \begin{bmatrix} 8 & 0 & 0 \\ 0 & 2/3 & -2 \\ 0 & 0 & 3 \end{bmatrix} \quad , \quad D = \begin{bmatrix} 6 & 0 & 3/2 \\ 0 & 2 & 1 \\ 0 & 0 & 7/2 \end{bmatrix}$$

a) Find a matrix B_1 *such that* $AB_1 = I$*. Show that we also have* $B_1A = I$*, so* $B_1 = A^{-1}$*.*

b) Find a matrix B_2 *such that* $CB_2 = I$*. Show that we also have* $B_2C = I$*, so* $B_2 = C^{-1}$*.*

c) Find a matrix B_3 *such that* $DB_3 = I$*. Show that we also have* $B_3D = I$*, so* $B_3 = D^{-1}$*.*

EXERCISE 2.68. *The following upper triangular matrices have entries in some field* $\mathbb{F}$*.*

$$A = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & 0 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 0 & a & b \\ 0 & c & d \\ 0 & 0 & e \end{bmatrix} \quad , \quad C = \begin{bmatrix} a & b & c \\ 0 & 0 & d \\ 0 & 0 & e \end{bmatrix}$$

a) Show that it is impossible for A *to be invertible.*

b) Show that it is impossible for B *to be invertible.*

c) Show that it is impossible for C *to be invertible.*

PART II.

Subspaces and linear independence

CHAPTER 3.

Subspaces of $\mathbb{F}^n$

Fix $\mathbb{F}$ as a field. Define $\mathbb{F}^n$ as the set of all n-tuples of elements of $\mathbb{F}$:

$$\mathbb{F}^n = \{(x_1, \ldots, x_n) : x_i \in \mathbb{F} \quad \forall i \in \{1, \ldots, n\}\}$$

Each element $x \in \mathbb{F}^n$ has the form $x = (x_1, ..., x_n)$. We call x a *vector*. We call $\mathbb{F}^n$ a *vector space*. The set $\mathbb{F}^n$ is in fact just one example of a more general definition of *vector space* that we develop in Chapter 6. This chapter focuses on $\mathbb{F}^n$.

We often write vectors in $\mathbb{F}^n$ as $x = (x_1, ..., x_n)$, without worrying about writing them vertically or horizontally. However, for purposes of arithmetic, each $x \in \mathbb{F}^n$ is viewed as a column vector (a $n \times 1$ matrix), so $x^\top$ is a row vector (a $1 \times n$ matrix). Given two vectors $x, y \in \mathbb{F}^n$ and a scalar $a \in \mathbb{F}$ we have by standard rules of matrix arithmetic:

$$x + y = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} + \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 + y_1 \\ \vdots \\ x_n + y_n \end{bmatrix}$$

$$ax = a \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} ax_1 \\ \vdots \\ ax_n \end{bmatrix}$$

Clearly $x + y = y + x$ and $a(x + y) = ax + ay$, as in the standard rules for matrices.

3.1. Linear combinations

Let $\{y_1, \ldots, y_k\}$ be a list of vectors in $\mathbb{F}^n$. Here, k is a positive integer that determines the number of vectors in the list. The list is allowed to have the same vector repeated more than once. A vector $z \in \mathbb{F}^n$ is a *linear combination of* $\{y_1, \ldots, y_k\}$ if

$$z = \sum_{i=1}^{k} a_i y_i$$

for some scalars $a_i \in \mathbb{F}$ for $i \in \{1, \ldots, k\}$. A linear combination can be viewed as a weighted sum where the scalars a_i are the weights. The sum always has a finite number of terms. [For general fields $\mathbb{F}$, infinite sums of vectors in $\mathbb{F}^n$ are not defined, see also Exercise 1.24.] The all-zero vector is a linear combination of every collection of vectors $\{y_1, ..., y_k\}$ in $\mathbb{F}^n$ because

$$0 = \sum_{i=1}^{k} 0 y_i$$

where the 0 on the left-hand-side is the all-zero vector $0 \in \mathbb{F}^n$, while the zeros on the right-hand-side are the scalars $0 \in \mathbb{F}$.

DEFINITION 3.1. *Let $\{y_1, \ldots, y_k\}$ be a list of vectors in $\mathbb{F}^n$. The* span *of these vectors, written $Span\{y_1, \ldots, y_k\}$, is the set of all linear combinations of $\{y_1, \ldots, y_k\}$. Specifically*

$$Span\{y_1, ..., y_k\} = \left\{ \sum_{i=1}^{k} a_i y_i : a_i \in \mathbb{F} \quad \forall i \in \{1, ..., k\} \right\}$$

Given a $m \times n$ matrix A with entries in $\mathbb{F}$, and given $x \in \mathbb{F}^n$, the vector $Ax \in \mathbb{F}^m$ is in fact a linear combination of the columns of A:

$$Ax = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}}_{x} = x_1 \begin{bmatrix} | \\ c_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ c_2 \\ | \end{bmatrix} + \cdots + x_n \begin{bmatrix} | \\ c_n \\ | \end{bmatrix}$$

where $c_1, ..., c_n$ are the columns of A. Therefore, the equation $Ax = b$ can be solved if and only if b is a linear combination of the columns of A. That is,

$$(Ax = b \text{ has a solution } x) \iff b \in Span\{\text{Columns of } A\}$$

3.1.1. Technical detail

The set $Span\{y_1, ..., y_k\}$ does not depend on the ordering of the y_i vectors that are listed:

$$Span\{y_1, y_2, y_3\} = Span\{y_2, y_3, y_1\} = Span\{y_1, y_3, y_2\}$$

And so on. This is because every linear combination of $\{y_1, ..., y_k\}$ is a vector of the form $\sum_{i=1}^{k} a_i y_i$ for some scalars a_i. Addition is commutative, so we can sum $a_i y_i$ in any order.

3.1.2. Example questions

1. Consider $\mathbb{F} = \mathbb{R}$. Let $y_1, y_2, y_3, z_1, z_2 \in \mathbb{R}^3$ be

$$y_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, y_2 = \begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}, y_3 = \begin{bmatrix} 4 \\ 0 \\ -2 \end{bmatrix}, z_1 = \begin{bmatrix} -3/2 \\ 0 \\ -8 \end{bmatrix}, z_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

a) Is $z_1 \in Span\{y_1, y_2, y_3\}$?
b) Is $z_2 \in Span\{y_1, y_2, y_3\}$?
c) Is $0 \in Span\{y_1, y_2, y_3\}$? (where $0 \in \mathbb{R}^3$ is the all-zero vector)

Solution:
a) To see if $z_1 \in Span\{y_1, y_2, y_3\}$, we must either find real numbers a, b, c that satisfy

$$a\underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}}_{y_1} + b\underbrace{\begin{bmatrix} 3 \\ 0 \\ 2 \end{bmatrix}}_{y_2} + c\underbrace{\begin{bmatrix} 4 \\ 0 \\ -2 \end{bmatrix}}_{y_3} = \underbrace{\begin{bmatrix} -3/2 \\ 0 \\ -8 \end{bmatrix}}_{z_1}$$

or we must prove that no such scalars exist. Equivalently, we must determine whether or not the following matrix equation has at least one solution $(a, b, c) \in \mathbb{R}^3$:

$$\begin{bmatrix} 1 & 3 & 4 \\ 0 & 0 & 0 \\ 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} -3/2 \\ 0 \\ -8 \end{bmatrix} \tag{3.1}$$

The matrix equation is equivalent to the following *system of three equations*:

$$a + 3b + 4c = -3/2 \tag{3.2}$$

$$0 + 0 + 0 = 0 \tag{3.3}$$

$$2b - 2c = -8 \tag{3.4}$$

In particular, (a, b, c) solves the system (3.2)-(3.4) if and only if it solves the matrix equation (3.1). If we *assume* (a, b, c) satisfies (3.2)-(3.4), we can often solve for a simple form that (a, b, c) must take. Assuming (3.2)-(3.4) hold, the equation (3.4) implies

$$\boxed{b = c - 4}$$

and substituting this into (3.2) gives

$$a + 3(c - 4) + 4c = -3/2 \implies \boxed{a = -7c + 21/2}$$

We only need one solution. For simplicity, set $c = 0$. Then we must have $b = -4$ and $a = 21/2$. It can be verified that $(21/2, -4, 0)$ indeed satisfies (3.1). Thus, the answer to (a) is: Yes, $z_1 \in Span\{y_1, y_2, y_3\}$.

b) If $z_2 \in Span\{y_1, y_2, y_3\}$, there must be real numbers a, b, c that satisfy

$$\begin{bmatrix} 1 & 3 & 4 \\ 0 & 0 & 0 \\ 0 & 2 & -2 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}}_{z_2} \tag{3.5}$$

No such a, b, c exist: For all $a, b, c \in \mathbb{R}$, the output of the matrix multiplication on the left-hand-side of (3.5) is a column vector with a zero on the second row, while the vector z_2 on the right-hand-side has a 1 on the second row. So $z_2 \notin Span\{y_1, y_2, y_3\}$.

c) Yes, $0 \in Span\{y_1, y_2, y_3\}$ because $0 = 0y_1 + 0y_2 + 0y_3$.

2. Consider $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic. Let $y_1, y_2, y_3, z_1, z_2 \in \mathbb{F}_2^4$ be

$$y_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}, y_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}, y_3 = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, z_1 = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}, z_2 = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}$$

[Notice that these vectors must have binary components.]
a) Is $z_1 \in Span\{y_1, y_2, y_3\}$?

b) Is $z_2 \in Span\{y_1, y_2, y_3\}$?
c) Is $0 \in Span\{y_1, y_2, y_3\}$? (where $0 = (0,0,0,0)$).

Solution:
a) To determine if $z_1 \in Span\{y_1, y_2, y_3\}$ we try to find scalars $a, b, c \in \{0,1\}$ that, under mod 2 arithmetic, satisfy the matrix equation

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}}_{z_1}$$

There are only $2^3 = 8$ possible scalars $(a, b, c) \in \{0,1\}^3$ and so it is easy to test them all. We find that $(a, b, c) = (1,1,1)$ works:

$$\begin{aligned} 1+1+0 &= 0 \pmod 2 \\ 0+1+1 &= 0 \pmod 2 \\ 1+1+1 &= 1 \pmod 2 \\ 1+0+0 &= 1 \pmod 2 \end{aligned}$$

Thus, $z_1 \in Span\{y_1, y_2, y_3\}$. [Note: If we were searching over solutions in $\{0,1\}^{300}$ rather than solutions in $\{0,1\}^3$, there would be more possibilities than the estimated number of atoms in the universe. We need a more efficient method for finding solutions to large problems. Fortunately, the method of *Gaussian elimination*, to be described later, can quickly find solutions without testing all possibilities.]

b) To determine if $z_2 \in Span\{y_1, y_2, y_3\}$ we try to find scalars $a, b, c \in \{0,1\}$ to solve

$$\begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \underbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 0 \end{bmatrix}}_{z_2}$$

We could test all 8 possible choices. Instead, let us write this as a system of 4 equations

(where equality holds in mod 2):

$$a + b = 1 \tag{3.6}$$

$$b + c = 0 \tag{3.7}$$

$$a + b + c = 1 \tag{3.8}$$

$$a = 0 \tag{3.9}$$

Suppose $(a, b, c) \in \{0, 1\}^3$ solves the system (3.6)-(3.9). Equality (3.9) implies

$$\boxed{a = 0}$$

Substituting $a = 0$ into (3.6) gives

$$0 + b = 1 \implies \boxed{b = 1}$$

Substituting $b = 1$ into (3.7) implies

$$1 + c = 0 \implies \boxed{c = 1} \quad \text{[recall } 1 + 1 = 0 \text{ in mod 2 arithmetic]}$$

So the only possible solution is $(a, b, c) = (0, 1, 1)$. However, this solution fails the equality (3.8):

$$(a + b + c = 1) \implies 0 + \underbrace{1 + 1}_{0} = 1 \implies 0 = 1 \quad \text{[contradiction]}$$

So there is no solution. That is, $z_2 \notin Span\{y_1, y_2, y_3\}$.

c) Indeed $0 \in Span\{y_1, y_2, y_3\}$ because $0 = 0y_1 + 0y_2 + 0y_3$.

3.2. Subspace of $\mathbb{F}^n$

DEFINITION 3.2. *A set $S \subseteq \mathbb{F}^n$ is a* subspace of $\mathbb{F}^n$ *if it satisfies the following three requirements:*

1. $0 \in S$ *(where* 0 *is the all-zero vector).*
2. *If $x, y \in S$ then $x + y \in S$.*
3. *If $x \in S$ and $a \in \mathbb{F}$ then $ax \in S$.*

The second property ensures the sum of two vectors in S is again in S. The third property ensures that a scalar multiple of a vector in S is again in S. When it is clear that we are working in the vector space $\mathbb{F}^n$, we often use the term "subspace" to implicitly mean "subspace of $\mathbb{F}^n$."

LEMMA 3.1. *A set $S \subseteq \mathbb{F}^n$ is a subspace if and only if it is nonempty and for all $x, y \in S$ and all $a, b \in \mathbb{F}$ we have*

$$ax + by \in S$$

PROOF. Suppose S is a subspace. Then it is nonempty because $0 \in S$. Fix $x, y \in S$ and $a, b \in \mathbb{F}$. Since scalar multiples of vectors in S are again in S, we know

$$ax \in S \quad , \quad by \in S$$

Since the sum of two vectors in S is again in S, we conclude

$$ax + by \in S$$

Conversely, suppose S is a nonempty subset of $\mathbb{F}^n$ with the property that for all $x, y \in S$ and all $a, b \in \mathbb{F}$ we have $ax + by \in S$. We want to show S is a subspace. Since S is nonempty, there is *at least one* vector in S. Fix $x, y \in S$ (this allows the possibility $x = y$, which is the *only* possibility in the special case when S has *only one* vector). Using $a = 0$ and $b = 0$ gives

$$\underbrace{0x + 0y}_{0} \in S$$

and so $0 \in S$. Using $a = 1$ and $b = 1$ gives

$$\underbrace{1x + 1y}_{x+y} \in S$$

and so $x + y \in S$. Finally, using any scalar $a \in \mathbb{F}$ and $b = 0$ gives

$$\underbrace{ax + 0y}_{ax} \in S$$

so $ax \in S$. Thus, S is a subspace. □

The above lemma ensures that every linear combination of two vectors in a subspace S is again in S. The next lemma uses induction to show that the linear combination of any finite number of vectors in a subspace S is again in S.

LEMMA 3.2. *Suppose S is a subspace of $\mathbb{F}^n$. Then for all $k \in \mathbb{N}$, all $v_1, \ldots, v_k \in S$, and all scalars $a_1, \ldots, a_k \in \mathbb{F}$ we have*

$$\sum_{i=1}^{k} a_i v_i \in S$$

PROOF. We use induction on k. The definition of subspace ensures the result holds for

$k = 1$ (all scalar multiples of a single vector in S is again in S). The previous lemma ensures it holds for $k = 2$ (the linear combination of 2 vectors in S is again in S). Suppose it holds for some integer $k \geq 2$. That is, suppose a linear combination of any k vectors in S is again in S. We show it also holds for $k + 1$. Fix $v_1, ..., v_{k+1} \in S$, $a_1, ..., a_{k+1} \in \mathbb{F}$. We want to prove

$$\sum_{i=1}^{k+1} a_i v_i \in S \tag{3.10}$$

Define

$$w = \sum_{i=1}^{k} a_i v_i$$

Since w is a linear combination of k vectors in S, our induction assumption implies $w \in S$. Then

$$\begin{aligned} \sum_{i=1}^{k+1} a_i v_i &= w + a_{k+1} v_{k+1} \\ &= 1w + a_{k+1} v_{k+1} \end{aligned}$$

Notice that $1w + a_{k+1} v_{k+1}$ is a linear combination of *two* vectors in S, so it is again in S. This proves the desired result (3.10). □

3.2.1. The trivial subspace $S = \{0\}$

Let $0 \in \mathbb{F}^n$ denote the all-zero vector. The set $S = \{0\}$ is called the *trivial subspace* of $\mathbb{F}^n$. It satisfies the definition of a subspace because: (i) $0 \in S$; (ii) The sum of any two vectors in S must be $0 + 0 = 0$, which is indeed in S; (iii) A scalar multiple of any vector in S must be $a0$ (for some $a \in \mathbb{F}$), but $a0 = 0 \in S$. The trivial subspace is the only subspace of $\mathbb{F}^n$ that contains only one vector. A subspace is called *nontrivial* if it is not the trivial subspace.

3.2.2. Examples

1. **Claim:** The vector space $\mathbb{F}^n$ is a subspace.

   PROOF. Clearly $\mathbb{F}^n \subseteq \mathbb{F}^n$. To show $\mathbb{F}^n$ is a subspace, first note that $0 = (0, ..., 0) \in \mathbb{F}^n$ (so $\mathbb{F}^n$ is nonempty). Next, fix $x, y \in \mathbb{F}^n$ and $a, b \in \mathbb{F}$. We have

   $$x = (x_1, ..., x_n), y = (y_1, ..., y_n)$$

   so

   $$ax + by = (ax_1 + by_1, ax_2 + by_2, ..., ax_n + by_n) \in \mathbb{F}^n$$

   so $\mathbb{F}^n$ is a subspace. □

2. **Claim:** Let k be a positive integer and let $\{v_1, ..., v_k\}$ be vectors in $\mathbb{F}^n$. Then $Span\{v_1, ..., v_k\}$ is a subspace.

   PROOF. The set $Span\{v_1, ..., v_k\}$ is the set of all linear combinations of $\{v_1, ..., v_k\}$. Since the v_i vectors are in $\mathbb{F}^n$, all linear combinations of them are also in $\mathbb{F}^n$, so the span is indeed a subset of $\mathbb{F}^n$.
   We now show the three properties of a subspace hold for $Span\{v_1, ..., v_k\}$. First observe $0 \in Span\{v_1, ..., v_k\}$ because $0 = \sum_{i=1}^{k} 0v_i$. Next, suppose x, y are vectors in $Span\{v_1, ..., v_k\}$. We must show that $x + y \in Span\{v_1, ..., v_k\}$. Since x, y are both in $Span\{v_1, ..., v_k\}$ we know

   $$x = \sum_{i=1}^{k} a_i v_i \tag{3.11}$$

   $$y = \sum_{i=1}^{k} b_i v_i \tag{3.12}$$

   for some scalars $a_1, ..., a_k$ and $b_1, ..., b_k$. Summing (3.11) and (3.12) gives

   $$x + y = \sum_{i=1}^{k} (a_i v_i + b_i v_i) = \sum_{i=1}^{k} \underbrace{(a_i + b_i)}_{\text{scalar}} v_i$$

   and this is a linear combination of $\{v_1, ..., v_k\}$. So $x + y \in Span\{v_1, ..., v_k\}$. Finally, fix $\theta \in \mathbb{F}$. We must show $\theta x \in Span\{v_1, ..., v_k\}$. From (3.11) we obtain

   $$\theta x = \theta \sum_{i=1}^{k} a_i v_i = \sum_{i=1}^{k} \underbrace{(\theta a_i)}_{\text{scalar}} v_i$$

   which is indeed a linear combination of $\{v_1, ..., v_k\}$. □

3. **Claim:** The intersection of two subspaces S_1 and S_2 of $\mathbb{F}^n$ is again a subspace of $\mathbb{F}^n$.

   PROOF. Let S_1, S_2 be two subspaces of $\mathbb{F}^n$. Let $S_1 \cap S_2$ denote the intersection. Clearly $S_1 \cap S_2 \subseteq \mathbb{F}^n$. Since the all-zero vector 0 must be in both S_1 and S_2, we know $0 \in S_1 \cap S_2$. Let x, y be vectors in $S_1 \cap S_2$ and let a, b be scalars in $\mathbb{F}$. We want to show $ax + by \in S_1 \cap S_2$. We have

   - Since $x, y \in S_1$ and S_1 is a subspace, we know $ax + by \in S_1$.
   - Since $x, y \in S_2$ and S_2 is a subspace, we know $ax + by \in S_2$.

   So $ax + by$ is in both S_1 and S_2, meaning it is in the intersection $S_1 \cap S_2$.

   □

4. Fix $\mathbb{F}$ as a field. Let A be a $m \times n$ matrix with entries in $\mathbb{F}$. Fix $b \in \mathbb{F}^m$ and assume $b \neq 0$. The set $\{x \in \mathbb{F}^n : Ax = b\}$ is the set of all solutions to the equation $Ax = b$.

   Claim: $\{x \in \mathbb{F}^n : Ax = b\}$ is *not* a subspace. [Recall the assumption $b \neq 0$]

   PROOF. The all-zero vector 0 is not in the set $\{x \in \mathbb{F}^n : Ax = b\}$ because

   $$A0 = 0 \neq b$$

   and so $\{x \in \mathbb{F}^n : Ax = b\}$ is not a subspace. □

3.3. The four fundamental subspaces of a matrix A

Let $A = (a_{ij})$ be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Vectors $x \in \mathbb{F}^n$ are viewed as column vectors so that the matrix multiplication Ax makes sense. Note that

$$x \in \mathbb{F}^n \implies Ax \in \mathbb{F}^m$$

For this matrix A, define

- $Null(A) = \{x \in \mathbb{F}^n : Ax = 0\}$. The set $Null(A)$ is called the *null space of A*. It is the set of all vectors $x \in \mathbb{F}^n$ that are mapped to zero through the matrix A.
- $Col(A) = \{Ax \in \mathbb{F}^m : x \in \mathbb{F}^n\}$. That is, $Col(A)$ is the set of all linear combinations of the columns of A. The set $Col(A)$ is called the *column space* of A.
- $Row(A) = Col(A^\top)$. The set $Row(A)$ is called the *row space* of A. It is the set of all linear combinations of the columns of $A^\top$. Since columns of $A^\top$ are the rows of A, we can also think of $Row(A)$ as the set of all linear combinations of the rows of A. However, the rows of A are row vectors. Since it is easier to think of all vectors as column vectors, we take the rows and turn them into columns. This is the reason we define $Row(A) = Col(A^\top)$.
- $Null(A^\top) = \{y \in \mathbb{F}^m : A^\top y = 0\}$. This is the null space of $A^\top$.

The sets $Null(A), Col(A), Null(A^\top), Col(A^\top)$ are called the *four fundamental subspaces of matrix A*. (Of course, $Row(A)$ is the same thing as $Col(A^\top)$.) We now formally show that these are indeed subspaces.

LEMMA 3.3. *Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Then*

- *The set $Null(A)$ is a subspace of $\mathbb{F}^n$.*
- *The set $Col(A)$ is a subspace of $\mathbb{F}^m$.*

PROOF. First consider $Null(A)$. Clearly $Null(A) \subseteq \mathbb{F}^n$. Also, $0 \in Null(A)$ because $A0 = 0$.

Next, fix $x, y \in Null(A)$ and $a, b \in \mathbb{F}$. It suffices to show that $ax + by \in Null(A)$. We have

$$A(ax + by) = aAx + bAy = a0 + b0 = 0$$

where we have used $Ax = 0$ and $Ay = 0$ (because $x, y \in Null(A)$). Thus, $A(ax + by) = 0$, so $ax + by \in Null(A)$.

Now consider $Col(A)$. Clearly $Col(A) \subseteq \mathbb{F}^m$. First observe that $0 \in Col(A)$ because $A0 = 0$. Next, fix $x, y \in Col(A)$ and $a, b \in \mathbb{F}$. It suffices to show $ax + by \in Col(A)$. Since $x, y \in Col(A)$ we know

$$x = Au$$
$$y = Av$$

for some vectors $u, v \in \mathbb{F}^n$. Thus

$$ax + by = a(Au) + b(Av) = A(au + bv)$$

Thus, $ax + by$ is the output when the matrix A left-multiplies the vector $(au + bv)$, so $ax + by \in Col(A)$. □

Since the above lemma ensures $Null(A)$ and $Col(A)$ are both subspaces for any matrix A, we can apply that result to the matrix $A^\top$ to immediately conclude that $Null(A^\top)$ and $Col(A^\top)$ are subspaces. The next lemma considers $Null(A)$ and $Col(A)$ for the special case when A is a square and invertible matrix.

LEMMA 3.4. *Fix $\mathbb{F}$ as a field. Let A be $n \times n$ matrix with entries in $\mathbb{F}$. If A is an invertible matrix then*

$$Null(A) = \{0\}$$
$$Col(A) = \mathbb{F}^n$$

Since $Null(A)$ consists of the single vector 0, we say that A has the trivial null space.

PROOF. To show $Null(A) = \{0\}$, fix $x \in Null(A)$. Then

$$Ax = 0$$

Multiplying the above by A^{-1} gives

$$A^{-1}Ax = A^{-1}0$$

Using the fact $A^{-1}A = I$ and $A^{-1}0 = 0$ gives

$$Ix = 0$$

Since $Ix = x$, we conclude $x = 0$. We have shown that an arbitrarily chosen vector in $Null(A)$ must be the all-zero vector. Thus, the *only* element of $Null(A)$ is the all-zero vector.

We now show $Col(A) = \mathbb{F}^n$. By definition, $Col(A) \subseteq \mathbb{F}^n$. To show $Col(A)$ and $\mathbb{F}^n$ are the same set, it suffices to show $\mathbb{F}^n \subseteq Col(A)$. Fix $y \in \mathbb{F}^n$. We show there is a vector $x \in \mathbb{F}^n$ such that $Ax = y$. Define

$$x = A^{-1}y$$

This works:

$$Ax = AA^{-1}y = Iy = y$$

□

3.3.1. Terminology

- The null space of a matrix A is also called the *kernel* of A:[1]

$$Null(A) = ker(A)$$

- The column space of a matrix A can also be called the *image* of A, and in some texts is called the *range* of A:

$$Col(A) = Image(A) = Range(A)$$

3.4. Speed with definitions

Let A be a matrix with entries in some field. Let x, y, v, w be vectors with sizes that make sense for the context used.

Question: Write the immediate consequences of these statements:

1. $y \in Col(A)$.
2. $x \in Null(A)$.
3. $v \in Row(A)$.
4. $w \in Null(A^\top)$.

[1] One semester when Neely was teaching linear algebra, a teaching assistant included a homework problem that used the term "kernel" and the notation "$ker(A)$." That week, Neely received multiple requests for an extension of the homework due date because the mysterious topic of "kernel" had not yet been covered in the lecture. Neely told students that it was the same as $Null(A)$, which had been covered several weeks before.

Solution:

1. $y \in Col(A) \implies y = Az$ for some vector z.
2. $x \in Null(A) \implies Ax = 0$.
3. $v \in Row(A) \implies v \in Col(A^\top) \implies v = A^\top z$ for some vector z.
4. $w \in Null(A^\top) \implies A^\top w = 0$.

Note: The phrase "for some vector z" is crucial for good mathematical communication. It is good to use this phrase. "Speed" does not mean we leave out necessary information, or that we write things without understanding them. It means we calmly and confidently proceed from one step to the next using basic mathematical techniques. This is a form of Mathematical Jiu-Jitsu. We have used "z" for the particular vector that exists because the variable z was not used in the given information. This assumes that bullet points 1 and 3 are associated with different problems so there is no conflict with reusing "z." Similar questions are given in the worksheet of Appendix A.

3.5. Operations that do not change $Null(A)$ or $Row(A)$

Recall that two sets C and D are equal if and only if $C \subseteq D$ and $D \subseteq C$. Suppose we are given two sets C and D, we are also given some information about these sets, and then we are asked to prove $C = D$. We can use the following two step procedure:

1. Fix $x \in C$. Use the given information to prove $x \in D$. This means $C \subseteq D$.
2. Fix $x \in D$. Use the given information to prove $x \in C$. This means $D \subseteq C$.

By proving $C \subseteq D$ and $D \subseteq C$, we conclude $C = D$.

LEMMA 3.5. *Fix $\mathbb{F}$ as a field. Let A be a $m \times n$ matrix with entries in $\mathbb{F}$. Let G be an invertible $m \times m$ matrix (so the multiplication GA makes sense). Then*

$$Null(GA) = Null(A)$$
$$Row(GA) = Row(A)$$

That is, left-multiplying A by an invertible matrix does not change its null space or its row space.

PROOF. There are four parts. Each part can be done using the "speed" techniques of the previous section.

1. Here we show $Null(GA) \subseteq Null(A)$. Fix $x \in Null(GA)$. We want to show $x \in Null(A)$.

Since $x \in Null(GA)$ we know

$$GAx = 0$$

Multiplying both sides of this equation by G^{-1} gives

$$\underbrace{G^{-1}G}_{I}Ax = \underbrace{G^{-1}0}_{0}$$

Using the fact $G^{-1}G = I$ and $G^{-1}0 = 0$ gives

$$IAx = 0$$

Since $IA = A$, this means $Ax = 0$, so $x \in Null(A)$.

2. Here we show $Null(A) \subseteq Null(GA)$. Fix $x \in Null(A)$. Then

$$Ax = 0$$

Multiplying the above equation by G gives

$$GAx = G0 = 0$$

and so $x \in Null(GA)$.

Parts 1 and 2 prove $Null(GA) = Null(A)$.

3. Here we show $Row(GA) \subseteq Row(A)$. Fix $y \in Row(GA)$. We want to show $y \in Row(A)$. By definition of $Row(GA)$ we have

$$y \in Col((GA)^\top) = Col(A^\top G^\top)$$

By definition of column space, it follows that

$$y = A^\top G^\top w$$

for some vector w. But then

$$y = A^\top \underbrace{(G^\top w)}_{\text{another vector}}$$

so $y \in Col(A^\top) = Row(A)$.

4. Here we show $Row(A) \subseteq Row(GA)$. Fix $y \in Row(A)$. We want to show $y \in Row(GA)$.

Since $y \in Row(A)$ we have (by definition of $Row(A)$):

$$y \in Col(A^\top)$$

so

$$y = A^\top w$$

for some vector w. Since G is invertible, we know $G^\top$ is invertible (recall Lemma 2.7). Since $w = Iw$ we have

$$y = A^\top w = A^\top Iw = A^\top \underbrace{G^\top (G^\top)^{-1}}_{I} w = (GA)^\top \underbrace{(G^\top)^{-1} w}_{\text{some vector}}$$

and so

$$y \in Col((GA)^\top) = Row(GA)$$

Parts 3 and 4 prove $Row(GA) = Row(A)$.

□

The process of *Gaussian elimination*, discussed in Chapter 4, involves left-multiplication of a general $m \times n$ matrix A by a sequence of invertible matrices. Each such left-multiplication does not change $Null(A)$ or $Row(A)$. The following lemma involves *right-multiplication*. Its proof is similar and is left as an exercise (Exercise 3.33).

LEMMA 3.6. *Fix $\mathbb{F}$ as a field. Let A be a $m \times n$ matrix with entries in $\mathbb{F}$. Let G be an invertible $n \times n$ matrix (so the multiplication AG makes sense). Then*

$$Col(A) = Col(AG)$$

That is, right-multiplying A by an invertible matrix does not change its column space.

3.6. Counting the vectors in a subspace

Given a set D, let $|D|$ denote the number of elements in D. Recall that

$$\mathbb{F}^n = \{(x_1, \ldots, x_n) : x_i \in \mathbb{F} \quad \forall i \in \{1, \ldots, n\}\}$$

Since $(x, 0, 0, \ldots, 0) \in \mathbb{F}^n$ for each $x \in \mathbb{F}$, we have

$$(|\mathbb{F}| = \infty) \implies (|\mathbb{F}^n| = \infty)$$

If $\mathbb{F}$ is a finite field of size f then $|\mathbb{F}^n| = f^n$. This is because there are f^n ways to choose $(x_1, ..., x_n)$:

- There are f choices of x_1;
- There are f choices of x_2;
- There are f choices of x_3;
- ...
- There are f choices of x_n

So the total number of choices is

$$\underbrace{f \cdot f \cdot f \cdots f}_{n \text{ times}} = f^n \implies |\mathbb{F}^n| = f^n$$

For example

- $\mathbb{F}_2 = \{0, 1\} \implies |\mathbb{F}_2^n| = 2^n$.

$$\mathbb{F}_2^2 = \{(0,0), (0,1), (1,0), (1,1)\}$$
$$\mathbb{F}_2^3 = \{(0,0,0), (0,0,1), (0,1,0), (0,1,1), (1,0,0), (1,0,1), (1,1,0), (1,1,1)\}$$

- $\mathbb{F}_3 = \{0, 1, 2\} \implies |\mathbb{F}_3^n| = 3^n$.

$$\mathbb{F}_3^2 = \{(0,0), (0,1), (0,2), (1,0), (1,1), (1,2), (2,0), (2,1), (2,2)\}$$

3.6.1. Example: Counting the vectors in $Span\{y_1, y_2, y_3, y_4\}$

Let $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic. The set $\mathbb{F}_2^3$ has 8 vectors:

$$\mathbb{F}_2^3 = \left\{ \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \right\}$$

Define the following particular vectors in $\mathbb{F}_2^3$:

$$y_1 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}, y_2 = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, y_3 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, y_4 = \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

The set of all linear combinations of $\{y_1, y_2, y_3, y_4\}$ is

$$Span\{y_1, y_2, y_3, y_4\} = \{a_1y_1 + a_2y_2 + a_3y_3 + a_4y_4 : a_1, a_2, a_3, a_4 \in \{0, 1\}\}$$

where arithmetic is in the binary field. How many vectors are in $Span\{y_1, y_2, y_3, y_4\}$?

Since there are $2^4 = 16$ ways to choose the binary coefficients (a_1, a_2, a_3, a_4) in the linear combination $\sum_{i=1}^{4} a_i y_i$, we might guess $|Span\{y_1, y_2, y_3, y_4\}| = 16$. However,

$$Span\{y_1, y_2, y_3, y_4\} \subseteq \mathbb{F}_2^3$$

and $\mathbb{F}_2^3$ has only 8 vectors. It is *impossible* for $Span\{y_1, y_2, y_3, y_4\}$ to contain 16 vectors because there are not that many vectors in the vector space! This means we are *double-counting*: There must be two distinct coefficient tuples (a_1, a_2, a_3, a_4) and (b_1, b_2, b_3, b_4) that lead to the same linear combination:

$$\sum_{i=1}^{4} a_i y_i = \sum_{i=1}^{4} b_i y_i$$

This holds if and only if

$$\sum_{i=1}^{4} (a_i - b_i) y_i = 0 \tag{3.13}$$

where the 0 on the right-hand-side is the all-zero vector. Define

$$z = (a_1, a_2, a_3, a_4) - (b_1, b_2, b_3, b_4)$$

so $z_i = a_i - b_i$ for $i \in \{1, \ldots, 4\}$. With this notation equation (3.13) becomes

$$\sum_{i=1}^{4} z_i y_i = 0$$

This can be written in matrix form

$$\underbrace{\begin{bmatrix} 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 \end{bmatrix}}_{\text{call this } A} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \\ z_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.14}$$

Since (a_1, a_2, a_3, a_4) is not the same as (b_1, b_2, b_3, b_4), the vector $z = (z_1, z_2, z_3, z_4)$ is not the all-zero vector. This means our matrix equation $Az = 0$ in (3.14) has a nontrivial solution!

(Meaning a solution $z \neq 0$). One such solution is

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

(Remember we are working with mod 2 arithmetic.) One reason there is a nontrivial solution is that $y_4 = y_1$ (so the first and fourth columns of A are identical). The vector y_4 does not introduce anything new. It is not difficult to prove

$$Span\{y_1, y_2, y_3, y_4\} = Span\{y_1, y_2, y_3\}$$

so throwing away y_4 does not change the span. Then

$$Span\{y_1, y_2, y_3\} = \left\{ \sum_{i=1}^{3} a_i y_i : a_i \in \{0, 1\} \quad \forall i \in \{1, 2, 3\} \right\}$$

Since y_1, y_2, y_3 are now *distinct* vectors, we might guess each distinct choice of coefficients (a_1, a_2, a_3) leads to a distinct linear combination $\sum_{i=1}^{3} a_i y_i$. *But this is also false.* To find redundancies, consider

$$\left(\sum_{i=1}^{3} a_i y_i = \sum_{i=1}^{3} b_i y_i \right) \iff \sum_{i=1}^{3} \underbrace{(a_i - b_i)}_{\text{call this } z_i} y_i = 0$$

We can set up the homogeneous equation as before

$$\sum_{i=1}^{3} z_i y_i = 0 \iff \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \\ z_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

There is *still* a nontrivial solution: $z = (1, 1, 1)$. This is because $y_3 = y_1 + y_2$. It is not difficult to prove

$$Span\{y_1, y_2, y_3\} = Span\{y_1, y_2\}$$

so throwing away y_3 does not change the span. The homogeneous equation for testing

$\{y_1, y_2\}$ for redundancies is

$$\sum_{i=1}^{2} z_i y_i = 0 \iff \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} z_1 \\ z_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

It is not difficult to show the only solution is the trivial solution $z = (0, 0)$. There is no more double-counting. Our final answer is

$$|Span\{y_1, y_2, y_3, y_4\}| = |Span\{y_1, y_2\}| = 2^2 = 4$$

3.6.2. Summarizing questions

Suppose $\mathbb{F}$ is a finite field with f scalars. Let $y_1, ..., y_4$ be vectors in $\mathbb{F}^n$. Recall that

$$Span\{y_1, y_2, y_3, y_4\} = \left\{ \sum_{i=1}^{4} a_i y_i : a_i \in \mathbb{F} \quad \forall i \in \{1, ..., 4\} \right\}$$

1. **Multiple choice:**
   What can be said about the number of vectors in $Span\{y_1, y_2, y_3, y_4\}$?
   a) $1 \leq |Span\{y_1, y_2, y_3, y_4\}| \leq f^n$.
   b) $1 \leq |Span\{y_1, y_2, y_3, y_4\}| \leq f^4$.
   c) Both (a) and (b).
   d) Impossible to tell.

2. **Multiple choice:**
   What can be said about comparing $Span\{y_1, y_2, y_3\}$ and $Span\{y_1, y_2, y_3, y_4\}$?
   a) $Span\{y_1, y_2, y_3\} \subseteq Span\{y_1, y_2, y_3, y_4\}$.
   b) $Span\{y_1, y_2, y_3\} \supseteq Span\{y_1, y_2, y_3, y_4\}$.
   c) $Span\{y_1, y_2, y_3\} = Span\{y_1, y_2, y_3, y_4\}$.
   d) None of the above.

3. **True/False:**
   $Span\{y_1, y_2, y_3\} = Span\{y_1, y_2, y_3, y_3\}$

   Solutions:

1. (c). $Span\{y_1, y_2, y_3, y_4\}$ contains at least one vector (the all zero vector). There are f^n vectors in the vector space $\mathbb{F}^n$, so $Span\{y_1, y_2, y_3, y_4\}$ cannot have more vectors than this! So (a) is true. Also, there is at most one vector in $Span\{y_1, y_2, y_3, y_4\}$ per distinct 4-tuple of scalar coefficients (a_1, a_2, a_3, a_4). There are f^4 coefficient tuples, so

$|Span\{y_1, y_2, y_3, y_4\}| \leq f^4$ and (b) is also true.

2. (a). Every linear combination of $\{y_1, y_2, y_3\}$ is also a linear combination of $\{y_1, y_2, y_3, y_4\}$ because for any scalars a_1, a_2, a_3 we have

$$a_1y_1 + a_2y_2 + a_3y_3 = a_1y_1 + a_2y_2 + a_3y_3 + 0y_4$$

3. True. Every linear combination of $\{y_1, y_2, y_3, y_3\}$ is also a linear combination of $\{y_1, y_2, y_3\}$ because for any scalars a_1, a_2, a_3, a_4 we have

$$a_1y_1 + a_2y_2 + a_3y_3 + a_4y_3 = a_1y_1 + a_2y_2 + (a_3 + a_4)y_3$$

3.7. Linear independence

DEFINITION 3.3. *Fix $\mathbb{F}$ as a field. A collection of vectors $\{v_1, \ldots, v_k\}$ in $\mathbb{F}^n$ is* linearly independent *if the following holds: Whenever $a_1, \ldots, a_k$ are scalars that satisfy*

$$\sum_{i=1}^{k} a_i v_i = 0$$

then we must have $a_i = 0$ for all $i \in \{1, \ldots, k\}$. In other words, the only way to create a linear combination of $\{v_1, \ldots, v_k\}$ that gives the all-zero vector is to choose all scalar weights as zero.

An equivalent definition of linear independence is this: Vectors $\{v_1, ..., v_k\}$ are linearly independent if the equation

$$\sum_{i=1}^{k} a_i v_i = 0$$

has only the trivial solution $(a_1, ..., a_k) = (0, 0, ..., 0)$.

DEFINITION 3.4. *Vectors $\{v_1, \ldots, v_k\}$ in $\mathbb{F}^n$ are* linearly dependent *if they are not linearly independent.*

It follows that if vectors $\{x_1, \ldots, x_k\}$ are linearly *dependent*, then there must be some scalars $a_1, \ldots, a_k$ such that at least one of the scalars is nonzero, and such that $\sum_{i=1}^{k} a_i x_i = 0$.

LEMMA 3.7. *Let A be a $m \times n$ matrix with entries in some field. Then*

Columns of A are linearly independent

$\iff$ *$Null(A) = \{0\}$*

$\iff$ *The equation $Ax = 0$ has only the trivial solution $x = 0$*

PROOF. Suppose A has columns $c_1, ..., c_n$:

$$A = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}$$

The columns are linearly independent if and only if the equation

$$\sum_{i=1}^{n} a_i c_i = 0$$

has only the trivial solution $(a_1, ..., a_n) = (0, 0, ..., 0)$. This is equivalent to saying that the following matrix equation

$$\underbrace{\begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}}_{A} \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

has only the trivial solution $(a_1, ..., a_n) = (0, ..., 0)$. That is, $Null(A) = \{0\}$. This holds if and only if the equation $Ax = 0$ has only the trivial solution. □

LEMMA 3.8. *(Counting linear combinations of linearly independent vectors) Let $\{v_1, ..., v_k\}$ be vectors in $\mathbb{F}^n$ that are linearly independent. Then*

a) Every distinct choice of scalar coefficients $(a_1, ..., a_k)$ leads to a distinct linear combination

$$\sum_{i=1}^{k} a_i v_i$$

b) If $\mathbb{F}$ is a finite field with $|\mathbb{F}| = f$, the number of vectors in $Span\{v_1, ..., v_k\}$ is exactly f^k, that is

$$|Span\{v_1, ..., v_k\}| = f^k$$

PROOF. To prove (a), suppose there are distinct coefficient vectors $(a_1, ..., a_k)$ and $(b_1, ..., b_k)$ such that

$$\sum_{i=1}^{k} a_i v_i = \sum_{i=1}^{k} b_i v_i$$

(we reach a contradiction). Rearranging the above equation gives

$$\sum_{i=1}^{k}(a_i - b_i)v_i = 0 \tag{3.15}$$

Since $\{v_1, ..., v_k\}$ are linearly independent, equation (3.15) immediately implies

$$(a_i - b_i) = 0 \quad \forall i \in \{1, ..., k\}$$

which means $(a_1, ..., a_k) = (b_1, ..., b_k)$, contradicting the assumption that these coefficient vectors were distinct. This proves part (a). Part (b) follows from (a) because there are exactly f^k ways to choose the scalar coefficients $(a_1, ..., a_k)$. □

3.7.1. Grammatical convention

For linearly independent vectors $\{v_1, ..., v_k\}$, should we say "is" or "are"? The convention we follow is that a *collection* of vectors is treated singular, while the vectors themselves are plural. So a collection of vectors $\{v_1, ..., v_k\}$ *is* linearly independent if and only if the vectors $\{v_1, ..., v_k\}$ *are* linearly independent. Theorems that deal with a general linearly independent collection $\{v_1, ..., v_k\}$ typically allow both the case $k = 1$ and $k > 1$. [Section 3.7.3 affirms that a collection that contains only one vector is linearly independent if and only if that vector is nonzero.] In such cases, we still say the vectors $\{v_1, ..., v_k\}$ *are* linearly independent because we allow the *possibility* of more than one vector. Of course, usage of "is" versus "are" does not change the math.

3.7.2. Technical detail

Linear independence of a collection of vectors $\{v_1, ..., v_k\}$ does not depend on the ordering of the v_i vectors in this list. Specifically, if $\{v_1, v_2, v_3\}$ are linearly independent, then so are $\{v_2, v_3, v_1\}$ and $\{v_1, v_3, v_2\}$ and so on.

If a collection of vectors is linearly independent, then any nonempty subcollection of those vectors is also linearly independent. This is formally proven in the following lemma.

LEMMA 3.9. *Let $\{v_1, ..., v_k\}$ be linearly independent (where $k \geq 2$). Then $\{v_1, ..., v_{k-1}\}$ are also linearly independent.*

PROOF. We want to show $\{v_1, ..., v_{k-1}\}$ are linearly independent. Suppose $a_1, ..., a_{k-1}$ are scalars that satisfy

$$\sum_{i=1}^{k-1} a_i v_i = 0$$

We want to show $a_i = 0$ for all $i \in \{1, ..., k-1\}$. The above equation implies

$$a_1 v_1 + a_2 v_2 + \cdots + a_{k-1} v_{k-1} + 0 v_k = 0$$

and since $\{v_1, ..., v_k\}$ are linearly independent, the above equation immediately implies

$$a_1 = 0, \quad a_2 = 0, \quad ..., \quad a_{k-1} = 0, \quad 0 = 0$$

□

3.7.3. Speed with linear independence

Question: Write the immediate consequence of the given information.

1. $\{v_1, ..., v_5\}$ are linearly independent vectors and $a_1, ..., a_5$ are scalars that satisfy
$$\sum_{i=1}^{5} a_i v_i = 0$$
2. $\{x, y, z\}$ are linearly dependent vectors.
3. $\{x_1, ..., x_k, y_1, ..., y_m\}$ are linearly dependent vectors.
4. $\{x_1, ..., x_k, y_1, ..., y_m\}$ are linearly independent vectors and $a_1, ..., a_k, b_1, ..., b_m$ are scalars that satisfy
$$\sum_{i=1}^{k} a_i x_i + \sum_{j=1}^{m} b_j y_j = 0$$

Solution:

1. $a_i = 0$ for all $i \in \{1, ..., 5\}$.
2. $ax + by + cz = 0$ for some scalars a, b, c, not all zero.
3. $\sum_{i=1}^{k} a_i x_i + \sum_{j=1}^{m} b_j y_j = 0$ for some scalars $a_1, ..., a_k, b_1, ..., b_m$, not all zero.
4. $a_i = 0$ for all $i \in \{1, ..., k\}$ and $b_j = 0$ for all $j \in \{1, ..., m\}$.

Question: Let x, y, z be vectors in $\mathbb{F}^{100}$, where $\mathbb{F}$ is some field.

1. Are $\{x, y, z, 0\}$ linearly independent? (where 0 is the all-zero vector in $\mathbb{F}^{100}$)
2. Are $\{x, y, z, z\}$ linearly independent?

Solution:

1. No. For a nontrivial solution to $ax + by + cz + d0 = 0$, we can use $(a, b, c, d) = (0, 0, 0, 1)$ (recall that scalars 0 and 1 are always in field $\mathbb{F}$) so

$$(0)x + (0)y + (0)z + (1)0 = 0$$

   Thus, $\{x, y, z, 0\}$ are not linearly independent. In general, a collection of linearly independent vectors cannot contain the all-zero vector.

2. No. For a nontrivial solution to $ax + by + cz + dz = 0$, we can use $(a, b, c, d) = (0, 0, 1, -1)$ (recall that scalars $0, 1, -1$ are always in field $\mathbb{F}$) so

$$(0)x + (0)y + (1)z + (-1)z = (1 + -1)z = 0z = 0$$

   Thus, $\{x, y, z, z\}$ are not linearly independent. In general, a collection of linearly independent vectors cannot contain a redundantly repeated vector.

Question: Let z be a nonzero vector in $\mathbb{F}^n$. Is $\{z\}$ linearly independent? That is, is a collection that consists of just one nonzero vector considered to be linearly independent?

Solution: Yes. To see why, let z be a nonzero vector in $\mathbb{F}^n$. Suppose c is a scalar such that

$$cz = 0 \tag{3.16}$$

We want to show $c = 0$. Suppose $c \neq 0$ (we reach a contradiction). If $c \neq 0$, the multiplicative inverse axiom ensures c^{-1} exists. Multiplying (3.16) by c^{-1} gives

$$c^{-1}cz = c^{-1}0$$

But $c^{-1}c = 1$ and $c^{-1}0 = 0$ so

$$1 \cdot z = 0$$

That is, $z = 0$, contradicting our assumption that $z \neq 0$.

3.8. Spanning a subspace

DEFINITION 3.5. *Let S be a subspace of $\mathbb{F}^n$. A collection of vectors $\{v_1, ..., v_k\}$ are said to* span *S if*

$$Span\{v_1, ..., v_k\} = S$$

Such vectors $\{v_1, ..., v_k\}$ *are also called a* spanning set *for subspace* S.

Notice that all vectors $\{v_1, ..., v_k\}$ in a spanning set for S must themselves be in S (since, for example, $v_1 = 1 \cdot v_1 + \sum_{i=2}^{k} 0 v_i$).

LEMMA 3.10. *A collection of vectors* $\{v_1, ..., v_k\}$ *span a subspace* S *if and only if all the vectors are in* S *and* $S \subseteq Span\{v_1, ..., v_k\}$.

PROOF. The forward direction is immediate. For the reverse direction, suppose $v_1, ..., v_k$ are all in S and $S \subseteq Span\{v_1, ..., v_k\}$. Since the v_i vectors are all in S, Lemma 3.2 ensures $Span\{v_1, ..., v_k\} \subseteq S$. Since we are also assuming $S \subseteq Span\{v_1, ..., v_k\}$, we conclude $S = Span\{v_1, ..., v_k\}$. □

Lemma 3.10 implies that to show a collection of vectors $\{v_1, ..., v_k\}$ span a subspace S, it is enough to show $v_i \in S$ for all $i \in \{1, ..., k\}$, and that all vectors in S can be written as a linear combination of $\{v_1, ..., v_k\}$.

3.9. Basis of a subspace

DEFINITION 3.6. *Let* S *be a subspace of* $\mathbb{F}^n$. *A collection of vectors* $\{v_1, ..., v_k\}$ *in* S *is called a* basis for S *if the following two properties hold*

1. $Span\{v_1, ..., v_k\} = S$.
2. $\{v_1, ..., v_k\}$ *are linearly independent.*

With this definition, if S is a subspace of $\mathbb{F}^n$, and if S has a basis $\{v_1, ..., v_k\}$, it means that all vectors $s \in S$ can be *uniquely* written as a linear combination of the basis vectors:

$$s = \sum_{i=1}^{k} a_i v_i$$

so the scalars $(a_1, ..., a_k)$ are unique to s. A vector $\tilde{s} \in S$ that satisfies $s \neq \tilde{s}$ would need scalars $(\tilde{a}_1, ..., \tilde{a}_k)$ such that

$$(a_1, ..., a_k) \neq (\tilde{a}_1, ..., \tilde{a}_k)$$

so $a_i \neq \tilde{a}_i$ for *at least one* $i \in \{1, ..., k\}$.

3.9.1. Basis for $\mathbb{F}^n$

Let $\{e_1, ..., e_n\}$ be the columns of the $n \times n$ identity matrix. The vectors e_i are in $\mathbb{F}^n$.

LEMMA 3.11. *The vectors* $\{e_1, ..., e_n\}$ *form a basis for the vector space* $\mathbb{F}^n$. *They are called the* standard basis for $\mathbb{F}^n$.

PROOF. We first show $\{e_1, ..., e_n\}$ span $\mathbb{F}^n$. Since $e_i \in \mathbb{F}^n$ for all i, it is enough to show that all vectors in $\mathbb{F}^n$ are linear combinations of $\{e_1, ..., e_n\}$ (recall Lemma 3.10). Let $y = (y_1, ..., y_n)$ be a vector in $\mathbb{F}^n$. Then

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = y_1 \underbrace{\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{e_1} + y_2 \underbrace{\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}}_{e_2} + \cdots + y_n \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}}_{e_n} = \sum_{i=1}^{n} y_i e_i$$

and so y is a linear combination of the vectors $\{e_1, ..., e_n\}$. This holds for all $y \in \mathbb{F}^n$, so $\mathbb{F}^n = Span\{e_1, ..., e_n\}$.

We now show $\{e_1, ..., e_n\}$ are linearly independent. Suppose there are scalars $a_1, ..., a_n$ such that

$$\sum_{i=1}^{n} a_i e_i = 0 \tag{3.17}$$

We want to show $a_i = 0$ for all i. The equation (3.17) is equivalent to

$$a_1 \underbrace{\begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}}_{e_1} + a_2 \underbrace{\begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \end{bmatrix}}_{e_2} + \cdots + a_n \underbrace{\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{bmatrix}}_{e_n} = 0$$

which can be written in matrix form as

$$\underbrace{Ia}_{a} = 0$$

where I is the $n \times n$ identity matrix and $a = (a_1, ..., a_n)$. Therefore $a = 0$. All scalars a_i are zero. □

3.9.2. Columns of a matrix

Let $\mathbb{F}$ be a field and let A be a $m \times n$ matrix with entries in $\mathbb{F}$. There are n columns of A:

$$A = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}$$

Let $\{c_1, ..., c_n\}$ be the columns of A. Then c_i is a column vector in $\mathbb{F}^m$ for all $i \in \{1, ..., n\}$. Recall that for every $x = (x_1, ..., x_n) \in \mathbb{F}^n$ we have

$$Ax = x_1 \begin{bmatrix} | \\ c_1 \\ | \end{bmatrix} + x_2 \begin{bmatrix} | \\ c_2 \\ | \end{bmatrix} + \cdots + x_n \begin{bmatrix} | \\ c_n \\ | \end{bmatrix}$$

and so Ax is a linear combination of the columns of A. Thus, the column space of A is the span of the columns of A:

$$Col(A) = Span\{c_1, ..., c_n\}$$

If the columns of A are linearly independent, then they form a basis for the subspace $Col(A)$. In that case, every vector in $Col(A)$ can be *uniquely* written as a linear combination of the columns of A (using unique scalar weights). In other words, every vector $y \in Col(A)$ has a *unique* vector $x \in \mathbb{F}^n$ such that $y = Ax$.

We summarize these findings in a lemma:

LEMMA 3.12. *Let $\mathbb{F}$ be a field and let A be a $m \times n$ matrix with entries in $\mathbb{F}$. Then*

a) $Col(A)$ is the span of the columns of A.

b) If the columns of A are linearly independent then:

- *The columns of A form a basis for $Col(A)$.*
- *For every $y \in Col(A)$, there is a unique $x \in \mathbb{F}^n$ such that $Ax = y$.*
- *$Null(A) = \{0\}$ (so $Ax = 0$ has only the trivial solution.)*

We omit a formal proof of this lemma since the first two bullet points have essentially been proven in the discussion immediately before the lemma. The last bullet point is just a repeat of the already-proven Lemma 3.7.

3.9.3. Dimension theory: Special case of finite fields

This subsection reaches several staggering conclusions about $\mathbb{F}^n$ by using simple *counting arguments*. The arguments only work when $\mathbb{F}$ is a finite field. However, the conclusions turn out to be true more generally for *any field* (which is shown in later chapters).

Let $\mathbb{F}$ be a finite field with f scalars (where $f \geq 2$). Let S be a subspace of $\mathbb{F}^n$. Suppose we have vectors $\{v_1, ..., v_k\}$ that form a basis for S. Since the basis vectors span S, for each $s \in S$ we have

$$s = \sum_{i=1}^{k} a_i v_i \tag{3.18}$$

for some scalars $a_1, ..., a_k$. The basis vectors are linearly independent, so the scalar coefficients $(a_1, ..., a_k)$ in (3.18) are *unique*. There are f^k tuples $(a_1, ..., a_k)$ and so

$$|S| = f^k \tag{3.19}$$

The *same* subspace S can have many different bases. Let $\{u_1, ..., u_m\}$ be some other basis for S, where m is some positive integer. We allow the possibility $m \neq k$ (perhaps this new basis for S has a different number of vectors than the first basis). Since $\{u_1, ..., u_m\}$ form a basis for S, and our field $\mathbb{F}$ has exactly f scalars, by the same reasoning used above we know

$$|S| = f^m \tag{3.20}$$

But S is the *same subspace* for both equations (3.19) and (3.20), so $f^k = f^m$ and $k = m$.

Conclusion 1: *Any two bases for the same subspace of $\mathbb{F}^n$ must have the same number of basis vectors.*

We can say more. Since S has k basis vectors, we know $|S| = f^k$. On the other hand, $S \subseteq \mathbb{F}^n$. We already know $\mathbb{F}^n$ has exactly f^n vectors. Thus

$$(S \subseteq \mathbb{F}^n) \implies (|S| \leq |\mathbb{F}^n|) \implies (f^k \leq f^n)$$

and we conclude $k \leq n$.

Conclusion 2: *If S is a subspace of $\mathbb{F}^n$, and if S has a basis, then the number of basis vectors for S is less than or equal to n.*

Similar counting arguments lead to the following theorem.

THEOREM 3.1. *(Wide matrix theorem – Special case of finite field) Let $\mathbb{F}$ be a finite field and let A be a $m \times n$ matrix with entries in this field. If A is a wide matrix (meaning it has more columns than rows) then it has a nontrivial null space. That is, $Null(A)$ contains a nonzero vector.*

For this proof we use the logical equivalence

$$(p \implies q) \equiv ((\text{not } q) \implies (\text{not } p))$$

where

- p = (Matrix A is wide)
- q = ($Null(A) \neq \{0\}$)
- (not p) = (Matrix A is not wide)
- (not q) = ($Null(A) = \{0\}$)

PROOF. Let $|\mathbb{F}| = f$. Suppose A has size $m \times n$. So it has m rows and n columns.

$$A = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}$$

Suppose $Null(A) = \{0\}$ (we shall show matrix A is not wide). Since $Null(A) = \{0\}$, Lemma 3.7 implies the columns of A are linearly independent. Hence, the n columns of A form a basis for $Col(A)$ and so

$$|Col(A)| = f^n$$

On the other hand, the columns of A are vectors in $\mathbb{F}^m$ so

$$Col(A) \subseteq \mathbb{F}^m \implies |Col(A)| \leq |\mathbb{F}^m|$$

Since $|Col(A)| = f^n$ and $|\mathbb{F}^m| = f^m$ we have

$$f^n \leq f^m$$

and so $n \leq m$. The number of columns is less than or equal to the number of rows, so the matrix is not wide. □

These counting arguments for finite fields provide good intuition. However, it is important to emphasize these remaining questions:

- Does every subspace of $\mathbb{F}^n$ have a basis?
- Do the above results hold for all fields $\mathbb{F}$ (not just for finite fields)?

3.9.4. The trivial subspace has no basis

Let $\mathbb{F}$ be any field. Define $S = \{0\}$, where 0 is the all-zero vector in $\mathbb{F}^n$. The set S is called the *trivial subspace* of $\mathbb{F}^n$. It is the only subspace that does not have a basis. To see why there is no basis, suppose $\{v_1, ..., v_k\}$ is a basis for S (we reach a contradiction). Since all basis vectors must be in S themselves, but $S = \{0\}$, it follows that $v_i = 0$ for all $i \in \{1, ..., k\}$. This contradicts the fact that a linearly independent collection of vectors cannot contain the all-zero vector (recall the question about $\{x, y, z, 0\}$ in Chapter 3.7.3).

The trivial subspace is considered to have *dimension* 0. Chapter 5 shows that all nontrivial subspaces of $\mathbb{F}^n$ have a basis, all bases of the same nontrivial subspace have the same number of vectors (defined as the *dimension* of the subspace), and the dimension of any nontrivial subspace of $\mathbb{F}^n$ is an integer in the set $\{1, ..., n\}$.

3.10. Proving linear independence: A case study

Suppose we are given a collection of vectors $\{x_1, \ldots, x_k\}$ that satisfy certain properties. We are asked to show the vectors are linearly independent. The first line of the proof is almost always the same:

(3.21)
Suppose $\sum_{i=1}^{k} a_i x_i = 0$ *for some scalars* $a_1, ..., a_k$. *We want to show* $a_i = 0$ *for all* $i \in \{1, ..., k\}$.

In other words, to prove $\{x_1, \ldots, x_k\}$ are linearly independent, we must show they satisfy the definition of linearly independent. The first line (3.21) instantiates a collection of scalars $a_1, \ldots, a_k$ for which we know $\sum_{i=1}^{k} a_i x_i = 0$ is a true equation. These scalars are held fixed throughout the proof (we are not allowed to choose them or change them). This sets us up for success. We then use the given properties about our vectors, together with our equation $\sum_{i=1}^{k} a_i x_i = 0$, to reach the goal of proving all a_i must be zero. Often this is accomplished just by performing the same operation on both sides of $\sum_{i=1}^{k} a_i x_i = 0$. This step-by-step method allows proofs of complicated statements to be done calmly and confidently, even if we do not have intuition about those statements.

Example problem: Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Let $\{y_1, \ldots, y_k\}$ be linearly independent vectors in $\mathbb{F}^m$. Let $\{x_1, \ldots, x_k\}$ be vectors in $\mathbb{F}^n$ that satisfy $Ax_i = y_i$ for all $i \in \{1, \ldots, k\}$. Prove that $\{x_1, \ldots, x_k\}$ are linearly independent.

Correct proof: Suppose that

(3.22)
$$\sum_{i=1}^{k} a_i x_i = 0$$

for some scalars $a_1, \ldots, a_k$. We want to show $a_i = 0$ for all i. Multiplying both sides of the equation (3.22) by A gives

$$\sum_{i=1}^{k} a_i A x_i = A0$$

Substituting $y_i = Ax_i$ and $A0 = 0$ gives

$$\sum_{i=1}^{k} a_i y_i = 0$$

which implies $a_i = 0$ for all i (because we are told $\{y_1, \ldots, y_k\}$ are linearly independent). □

Notice that the first line of this proof is the same as the suggested first line in (3.21). Of course, if we were trying to prove some other vectors $\{w_1, \ldots, w_5\}$ were linearly independent, we would need to modify (3.21) appropriately (change x_i to w_i and change k to 5). The two conditions given in this example were: (i) $\{y_1, \ldots, y_k\}$ are linearly independent; (ii) $Ax_i = y_i$ for all i. Even if we do not have any intuition about these two conditions, our correct first line, which establishes the equation (3.22), makes it easy to arrive at our goal. That is because we are given a fact about the vectors y_i, and we are also told that $Ax_i = y_i$. The x_i vectors in equation (3.22) *want* to be multiplied by A. We give them what they want by multiplying both sides of the equation (3.22) by A.

In summary, even though the two conditions in the example seem complicated, it is impossible *not* to prove the above result. Unfortunately, students often do the impossible. Common incorrect proofs are given below (see also Exercises 3.22, 3.23).

1. Most common incorrect start: "Suppose $\sum_{i=1}^{k} a_i y_i = 0$."
   [We are not trying to prove that $\{y_1, \ldots, y_k\}$ are linearly independent.]

2. Another incorrect proof: Let $\sum_{i=1}^{k} 0x_i = 0$. Choose $a_i = 0$. Then $\sum_{i=1}^{k} a_i x_i = 0$ and all $a_i = 0$, so $\{x_1, \ldots, x_k\}$ are linearly independent.
   [If this proof method worked, it would prove that *any* collection of vectors are linearly independent, which is absurd. To prove linear independence, we are not allowed to choose the a_i coefficients to be anything we want.]

3. Another incorrect proof: We know $\sum_{i=1}^{k} a_i y_i = 0$. Then $a_i = 0$ since $\{y_1, \ldots, y_k\}$ are linearly independent. Substitute the definition $y_i = Ax_i$ to get $\sum_{i=1}^{k} a_i A x_i = 0$. Then $A(\sum_{i=1}^{k} a_i x_i) = 0$. Multiply by A^{-1} to get $\sum_{i=1}^{k} a_i x_i = 0$. But we already know $a_i = 0$ for all i, so we are done.

[Even if A were invertible, this proof would not work. The first line "We know $\sum_{i=1}^{k} a_i y_i = 0$" has no meaning because no a_i coefficients have been defined. Even if we change the first line to something meaningful such as "Suppose $\sum_{i=1}^{k} a_i y_i = 0$ for some scalars a_i," then this does not help: Linear independence of $\{y_1, ..., y_k\}$ immediately implies $a_i = 0$ for all i. So, at best, this is similar to starting off by "assuming" the trivially true fact $\sum_{i=1}^{k} 0 y_i = 0$.]

4. Another incorrect approach: The student diligently writes the correct first line, which instantiates scalars $a_1, ..., a_k$ that satisfy equation (3.22). Then, the student *ignores this equation and redefines the scalars a_i as some other values* in a scattering of unusual statements that lead nowhere and/or that recreates one of the above incorrect proof instances.

5. Another incorrect proof: We know $\{x_1, \ldots, x_k\}$ are linearly independent because, basically, none of the vectors is a linear combination of the others, since we are told $Ax_i = y_i$ and also we know $\{y_1, ..., y_k\}$ are linearly independent.
   [This just restates the given information with no effort to show how that information can be used to reach a conclusion. On the positive side, the first sentence does indeed state a useful property related to linear independence.]

The most promising incorrect proof is the last one. Let us reformulate it into a correct proof. It is true that $\{x_1, ..., x_k\}$ are linearly *dependent* if and only if one of the x_i vectors is a linear combination of the others (this is proven in Section 5.1). Let us start with that.

Correct proof that does not use the suggested first line: Suppose $\{x_1, ..., x_k\}$ are *not* linearly independent (we reach a contradiction). Then one of the x_i vectors is a linear combination of the others. Without loss of generality suppose x_k is a linear combination of the others (if this is not true, just reorder the vectors). So

$$x_k = \sum_{i=1}^{k-1} a_i x_i$$

for some scalars a_i. Multiplying this equation by A gives

$$Ax_k = \sum_{i=1}^{k-1} a_i A x_i$$

Since $Ax_i = y_i$ for all i, the above equation reduces to

$$y_k = \sum_{i=1}^{k-1} a_i y_i$$

so y_k is a linear combination of $\{y_1, ..., y_{k-1}\}$, contradicting linear independence of $\{y_1, ..., y_k\}$. □

3.10.1. Another perspective

Given vectors $\{x_1, \ldots, x_k\}$, we can define S as the set of all $(a_1, ..., a_k) \in \mathbb{F}^k$ that satisfy

$$\sum_{i=1}^{k} a_i x_i = 0$$

Clearly S is nonempty because the all-zero vector is in S. The vectors $\{x_1, \ldots, x_k\}$ are linearly independent if and only if 0 is the *only* element of set S. Thus, to prove $\{x_1, ..., x_k\}$ are linearly independent, the following first lines of the proof are equivalent:

1. "Suppose $\sum_{i=1}^{k} a_i x_i = 0$ for some scalars a_i. We want to show $a_i = 0$ for all i."
2. "Fix $a \in S$. We want to show $a = 0$."

When we say "Fix $a \in S$," the understanding is that there is nothing special about the particular a that has been selected from S. If we can prove *that particular* a must be the all-zero vector, then the same proof can be repeated for *any other* $a \in S$ *that we could have considered*. If every $a \in S$ is the all-zero vector, then $S = \{0\}$.

3.11. Example: Establishing linear independence for two or three vectors

This example is similar to Exercises 3.19 and 3.20. Consider the following vectors in $\mathbb{R}^4$:

$$w = \begin{bmatrix} 10 \\ 0 \\ -1 \\ 2 \end{bmatrix}, \quad y = \begin{bmatrix} 2 \\ 33 \\ 5 \\ -5 \end{bmatrix}, \quad z = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix} \tag{3.23}$$

Claim 1: $\{w, y\}$ are linearly independent.

PROOF. (approach 1 to claim 1) Suppose $aw + by = 0$ for some scalars $a, b \in \mathbb{R}$. We show $a = b = 0$. [Observe this is the suggested first line (3.21), appropriately modified with $k = 2$, $x_1 = w, x_2 = y$.]

The equation $aw + by = 0$ can be written in matrix form:

$$\begin{bmatrix} 10 & 2 \\ 0 & 33 \\ -1 & 5 \\ 2 & -5 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$$

The matrix on the left-hand-side of the above equation has columns w and y from (3.23). This is equivalent to the following system of four equations

$$10a + 2b = 0 \tag{3.24}$$

$$33b = 0 \tag{3.25}$$

$$-a + 5b = 0 \tag{3.26}$$

$$2a - 5b = 0 \tag{3.27}$$

Equation (3.25) implies $b = 0$. Substituting $b = 0$ into (3.24) gives

$$10a + 0 = 0 \implies a = 0$$

and so we have proven $a = b = 0$. We are done. □

PROOF. (approach 2 to claim 1) Suppose $aw + by = 0$ for some scalars $a, b \in \mathbb{R}$. We show $a = b = 0$.

- Case 1: Suppose $a = 0$. Then $by = 0$. Since y is not the all-zero vector, we must have $b = 0$ and we are done.
- Case 2: Suppose $a \neq 0$. The equation $aw + by = 0$ reduces to $w = (-b/a)y$, that is

$$\underbrace{\begin{bmatrix} 10 \\ 0 \\ -1 \\ 2 \end{bmatrix}}_{w} = (-b/a) \underbrace{\begin{bmatrix} 2 \\ 33 \\ 5 \\ -5 \end{bmatrix}}_{y}$$

  Just from the first two rows, we see this means

$$10 = 2(-b/a)$$
$$0 = 33(-b/a)$$

The second equation implies $-(b/a) = 0$, and substituting that into the first equation gives $10 = 0$, a contradiction. So this case 2 cannot occur. The only relevant case is Case 1, and we are done.

□

Claim 2: The vectors $\{w, y, z\}$ in (3.23) are linearly independent.

PROOF. (approach 1 to claim 2) Suppose $aw + by + cz = 0$ for some scalars a, b, c. We want to show $a = b = c = 0$. The equation $aw + by + cz = 0$ can be written in matrix form:

$$\begin{bmatrix} 10 & 2 & 1 \\ 0 & 33 & 2 \\ -1 & 5 & 3 \\ 2 & -5 & 4 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} \tag{3.28}$$

The columns of the matrix on the left-hand-side are the vectors w, y, z from (3.23). We need some method of proving that $a = b = c = 0$.

The method of *Gaussian elimination* is developed in the next chapter (Chapter 4). Here we showcase the method. Readers who are not yet familiar with matrix row reductions of Gaussian elimination can skip the remainder of this example (they may want to work out the system of four equations and three unknowns for (3.28) using basic algebra).

Using Gaussian elimination to find all (a, b, c) solutions to (3.28) gives

$$\begin{bmatrix} 10 & 2 & 1 \\ 0 & 33 & 2 \\ -1 & 5 & 3 \\ 2 & -5 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 10 & 2 & 1 \\ 0 & 33 & 2 \\ 0 & 52 & 31 \\ 0 & 27 & -19 \end{bmatrix} \rightarrow \begin{bmatrix} 10 & 2 & 1 \\ 0 & 33 & 2 \\ 0 & 0 & 31 + (2)\frac{-52}{33} \\ 0 & 0 & -19 + (2)\frac{-27}{33} \end{bmatrix} \rightarrow \begin{bmatrix} \boxed{10} & 2 & 1 \\ 0 & \boxed{33} & 2 \\ 0 & 0 & \boxed{\frac{919}{33}} \\ 0 & 0 & 0 \end{bmatrix}$$

There are no free variables so the only solution is the trivial solution $(a, b, c) = (0, 0, 0)$. □

PROOF. (approach 2 to claim 2) Suppose $aw + by + cz = 0$ for some scalars a, b, c. We want to show $a = b = c = 0$.

- Case 1: Suppose $c = 0$. The equation $aw + by + cz = 0$ reduces to $aw + by = 0$. But claim 1 has already shown that $\{w, y\}$ are linearly independent, so we immediately conclude $a = b = 0$. So we are done.

- Case 2: Suppose $c \neq 0$. The equation $aw + by + cz = 0$ reduces to $z = (-a/c)w + (-b/c)y$. Define scalars $\theta_1 = -a/c$ and $\theta_2 = -b/c$. We want to find all solutions (θ_1, θ_2) to $\theta_1 w + \theta_2 y = z$. In matrix form this is

$$\begin{bmatrix} 10 & 2 \\ 0 & 33 \\ -1 & 5 \\ 2 & -5 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

  Using Gaussian elimination to find all solutions $\theta \in \mathbb{R}^2$ to this inhomogeneous equation $A\theta = z$ gives

$$\left[\begin{array}{cc|c} 10 & 2 & 1 \\ 0 & 33 & 2 \\ -1 & 5 & 3 \\ 2 & -5 & 4 \end{array}\right] \rightarrow \left[\begin{array}{cc|c} 10 & 2 & 1 \\ 0 & 33 & 2 \\ 0 & 52 & 31 \\ 0 & 27 & -19 \end{array}\right] \rightarrow \left[\begin{array}{cc|c} 10 & 2 & 1 \\ 0 & 33 & 2 \\ 0 & 0 & 31 + (2)\frac{-52}{33} \\ 0 & 0 & -19 + (2)\frac{-27}{33} \end{array}\right] = \left[\begin{array}{cc|c} \boxed{10} & 2 & 1 \\ 0 & \boxed{33} & 2 \\ 0 & 0 & \frac{919}{33} \\ 0 & 0 & \frac{-227}{11} \end{array}\right]$$

  This means $0 = 919/33$, a contradiction. So this case 2 cannot occur.

The reader may notice a similarity between the Gaussian elimination used in approach 1 and approach 2 for this problem. The same numbers such as 919/33 appear. □

3.12. Exercises

EXERCISE 3.1. *(Linear combinations in $\mathbb{R}^2$) Define the following vectors in $\mathbb{R}^2$:*

$$x = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \quad y = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

a) Show that $Span\{x, y\} = \mathbb{R}^2$.

b) Show that $\{x, y\}$ are linearly independent. Hence, $\{x, y\}$ is a basis for $\mathbb{R}^2$.

EXERCISE 3.2. *Consider $\mathbb{F} = \mathbb{R}$. Show the following three vectors in $\mathbb{R}^3$ are linearly independent.*

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, y = \begin{bmatrix} 3 \\ 1 \\ 0 \end{bmatrix}, z = \begin{bmatrix} 5 \\ -4 \\ 1 \end{bmatrix}$$

EXERCISE 3.3. *Consider $\mathbb{F} = \mathbb{R}$. Show the following three vectors in $\mathbb{R}^4$ are linearly independent.*

$$x = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 0 \end{bmatrix}, y = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 0 \end{bmatrix}, z = \begin{bmatrix} 5 \\ -4 \\ 1 \\ 9 \end{bmatrix}$$

EXERCISE 3.4. *Consider the upper triangular 3×3 matrix A given below, with entries in some field $\mathbb{F}$. Assume that $a \neq 0, d \neq 0, f \neq 0$.*

$$A = \begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}$$

a) Show that the columns of A are linearly independent.

b) Find Null(A).

EXERCISE 3.5. *(Computing Null(A)) Using field $\mathbb{R}$:*

a) Find Null(A) for

$$A = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$$

b) Find Null(A) for

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 4 \end{bmatrix}$$

EXERCISE 3.6. *(Computing Col(A)) Using field $\mathbb{R}$:*

a) Find Col(A) for

$$A = \begin{bmatrix} 2 & 4 \\ 2 & 4 \end{bmatrix}$$

b) Find Col(A) for

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 4 \end{bmatrix}$$

EXERCISE 3.7. *Consider the following vectors $\{x, y, z\}$ in $\mathbb{R}^3$*

$$x = \begin{bmatrix} 3 \\ 1 \\ 10 \end{bmatrix}, y = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, z = \begin{bmatrix} 4 \\ 2 \\ 11 \end{bmatrix}$$

a) Are $\{x, y, z\}$ linearly independent?

b) Can you find a vector $w \notin Span\{x, y, z\}$?

EXERCISE 3.8. *(Is this a subspace?) Fix $\mathbb{F}$ as a field. Let a, b, c denote particular scalars in $\mathbb{F}$. Are the following subsets of $\mathbb{F}^3$ subspaces? Prove true, or give a counter-example.*

a) $S_1 = \{(x_1, x_2, x_3) \in \mathbb{F}^3 : x_1 = x_2\}$

b) $S_2 = \{(x_1, x_2, x_3) \in \mathbb{F}^3 : ax_1 + bx_2 + cx_3 = 0\}$

c) $S_3 = \{(x_1, x_2, x_3) \in \mathbb{F}^3 : x_1 x_2 = 0\}$

d) $S_4 = \{(x_1, x_2, x_3) \in \mathbb{F}^3 : x_1^2 = 0\}$

e) $S_5 = \{(x_1, x_2, x_3) \in \mathbb{F}^3 : x_1 + x_2 = 1\}$

EXERCISE 3.9. *(Subspace or not, depending on $\mathbb{F}$) Let $\mathbb{F}$ be a field. Consider*

$$S = \{(x, y, z) \in \mathbb{F}^3 : x^2 + y^2 = 0\}$$

Explain why S may or may not be a subspace, depending on $\mathbb{F}$. [KISS = Keep It Simple and Specific. Find one specific field for which S is a subspace, then another specific field for which it is not.]

EXERCISE 3.10. *We use a field $\mathbb{F}_p = \{0, 1, ..., p-1\}$ for some prime p and mod p arithmetic. Our matrix A has size 2000×1000 and has entries in $\mathbb{F}_p$. We know that $Null(A)$ has exactly 6859 vectors.*

a) What is p?

b) How many vectors are in a basis for $Null(A)$?

c) Are the columns of A linearly independent?

EXERCISE 3.11. *We use a field $\mathbb{F}_p = \{0, 1, ..., p-1\}$ for some prime p and mod p arithmetic. Our matrix A has size 500×500 and has entries in $\mathbb{F}_p$. We know that $Null(A)$ has exactly 10510100501 vectors.*

a) What is p?

b) How many vectors are in a basis for $Null(A)$?

c) Is A invertible?

EXERCISE 3.12. *(Multiple choice) We have two sets S_1, S_2 and we want to prove $S_1 \subseteq S_2$. What is a correct start to the proof?*

a) We want to show there exists an $x \in S_1$ such that $x \in S_2$.

b) Fix $x \in S_2$. We want to show $x \in S_1$.

c) Fix $x \in S_1$. We want to show $x \in S_2$.

d) None of the above.

EXERCISE 3.13. *Let A be a $m \times n$ matrix with columns $c_1, ..., c_n$. Show that*

$$Col(A) = Span\{c_1, ..., c_n\}$$

EXERCISE 3.14. *Let A be a $m \times n$ matrix with entries in some field.*

a) Let $\tilde{A}$ be a $m \times n$ matrix that is created by permuting the columns of A. Show that $Col(\tilde{A}) = Col(A)$. Thus, permuting the columns of a matrix does not change the column space. Give a specific example.

b) Suppose $n \geq 2$ and the last column of A is a linear combination of the previous $n-1$ columns. Let B be the $m \times (n-1)$ matrix formed by removing the last column of A. Show that $Col(B) = Col(A)$. Thus, removing linearly dependent columns does not change the column space. Give a specific example.

EXERCISE 3.15. *Let A be a $m \times n$ matrix with entries in some field.*

a) Let $\tilde{A}$ be a $m \times n$ matrix that is created by permuting the rows of A. Prove $Row(\tilde{A}) = Row(A)$ and $Null(\tilde{A}) = Null(A)$.

b) Suppose $m \geq 2$ and the last row of A is a linear combination of the previous $m-1$ rows. Let B be the $(m-1) \times n$ matrix formed by removing the last row of A. Show that $Row(B) = Row(A)$.

c) For the same matrix B of part (b), show that $Null(B) = Null(A)$. Thus, removing linearly dependent rows of a matrix does not change its row space or its null space.

EXERCISE 3.16. *(Matrices with linearly independent columns) Suppose C and D are matrices such that the multiplication CD makes sense. Suppose columns of CD are linearly independent. Prove the columns of D are linearly independent.*

EXERCISE 3.17. *Let $\mathbb{F}$ be a field. Suppose $A \sim m \times n$ has linearly independent columns and $B \sim m \times m$ is invertible. Show that the $m \times n$ matrix BA has linearly independent columns. That is, show the equation $BAx = 0$ has only the trivial solution.*

EXERCISE 3.18. *Let x, y, z be vectors in $\mathbb{R}^5$.*

a) Are $\{0, x, y, z\}$ linearly independent?

b) Are $\{x, y, z, 2x + y - z\}$ linearly independent?

EXERCISE 3.19. *Consider these vectors $u, v, w \in \mathbb{R}^3$:*

$$u = \begin{bmatrix} 1 \\ 3 \\ 5 \end{bmatrix}, \quad v = \begin{bmatrix} 0 \\ 0 \\ -1 \end{bmatrix}, \quad w = \begin{bmatrix} 2 \\ -1 \\ 1 \end{bmatrix}$$

a) Are $\{u, v\}$ linearly independent?

b) Are $\{u, v, w\}$ linearly independent?

c) Are $\{u, u + v, u + v + w\}$ linearly independent?

EXERCISE 3.20. *Design your own example similar to the examples in Section 3.11:*

a) Design an example to show that three vectors $\{u, v, w\}$ in $\mathbb{R}^4$ are linearly independent.

b) Design an example to show that three vectors $\{u, v, w\}$ in $\mathbb{R}^4$ are linearly dependent.

EXERCISE 3.21. *(Easy conditions for linear dependence) Let $\{x_1, ..., x_k\}$ be a collection of vectors in $\mathbb{F}^n$.*

a) The zero vector is in the collection: Show that if $x_i = 0$ for some $i \in \{1, ..., k\}$ then $\{x_1, ..., x_k\}$ are linearly dependent.

b) Two vectors in the collection are repeated: Show that $x_i = x_j$ for some distinct indices $i, j \in \{1, ..., k\}$ such that $i \neq j$, then $\{x_1, ..., x_k\}$ are linearly dependent.

EXERCISE 3.22. *(Multiple choice) Suppose we are told that vectors $x_1, ..., x_n$ have certain properties. We want to use these properties to show $\{x_1, ..., x_n\}$ are linearly independent. Regardless of these specific properties, what is the logically correct first sentence of our proof? State the correct choice below. Then, for each incorrect choice, explain why it is incorrect.*

a) We want to show $\sum_{i=1}^n a_i x_i = 0$.

b) We want to show $\sum_{i=1}^n a_i x_i = 0$ for all scalars a_i.

c) We want to show $\sum_{i=1}^n a_i x_i = 0$ for some scalars a_i.

d) We want to show that if scalars $a_1, ..., a_n$ satisfy $\sum_{i=1}^n a_i x_i = 0$, then $a_i = 0$ for all i.

e) We want to show that if $a_i = 0$ for all i, then $\sum_{i=1}^n a_i x_i = 0$.

EXERCISE 3.23. *(Multiple choice) We want to show vectors $\{x_1, ..., x_k, z_1, ..., z_m, w_1, ..., w_n\}$ are linearly independent. What is the logically correct first sentence? For each incorrect choice, write a sentence to explain why it is incorrect.*

a) We want to show $\sum_{i=1}^k a_i x_i + \sum_{j=1}^m b_j z_j + \sum_{r=1}^n c_r w_r = 0$.

b) Suppose $a_1, ..., a_k$ are scalars and $\sum_{i=1}^k a_i x_i = 0$. We want to show $a_i = 0$ for all i.

c) Suppose $a_1, ..., a_{k+m+n}$ are scalars and $\sum_{i=1}^{k+m+n} a_i x_i = 0$. We want to show $a_i = 0$ for all i.

d) Suppose $a_1, ..., a_k, b_1, ..., b_m, c_1, ..., c_n$ are scalars and $\sum_{i=1}^k a_i x_i + \sum_{j=1}^m b_j z_j + \sum_{r=1}^n c_r w_r = 0$. We want to show all of the scalars $a_1, ..., a_k, b_1, ..., b_m, c_1, ..., c_n$ are zero.

EXERCISE 3.24. *(Null(A) and Col(A) are subspaces) Let A be a matrix of size $m \times n$ with entries in some field $\mathbb{F}$.*

a) Show that Null(A) is a subspace of $\mathbb{F}^n$.

b) Show that Col(A) is a subspace of $\mathbb{F}^m$.

EXERCISE 3.25. *(Proving linear independence) Let $\mathbb{F}$ be a field. Suppose $\{x_1, x_2, ..., x_{20}\}$ are vectors in $\mathbb{F}^n$ and $\{y_1, y_2, ..., y_{20}\}$ are vectors in $\mathbb{F}^m$ (m and n are given positive integers). Suppose there are twenty matrices $A_1, A_2, \ldots, A_{20}$ with the property that for each $k, j \in \{1, \ldots, 20\}$ we have:*

$$A_k x_j = \begin{cases} 0 & \text{if } k \neq j \\ y_k & \text{if } k = j \end{cases}$$

Suppose $y_i \neq 0$ for $i \in \{1, ..., 20\}$. Prove that $\{x_1, x_2, ..., x_{20}\}$ are linearly independent.

EXERCISE 3.26. *(Linear independence with distinct eigenvalues) Assume A is a square matrix with entries in some field $\mathbb{F}$. A scalar $\lambda \in \mathbb{F}$ is an* eigenvalue *of A if there is a nonzero vector v*

such that $Av = \lambda v$. The vector v is called an eigenvector. *Suppose there are distinct eigenvalues λ_1, λ_2 with eigenvectors v_1, v_2, so $\lambda_1 \neq \lambda_2$ and $v_i \neq 0, Av_i = \lambda_i v_i$ for $i \in \{1, 2\}$. Show that $\{v_1, v_2\}$ are linearly independent. [Hint: Write the standard first sentence "Suppose a, b are scalars such that $av_1 + bv_2 = 0$." Then consider the cases $a = 0$ and $a \neq 0$.]*

EXERCISE 3.27. *(Linear independence of generalized eigenvectors) Assume A is a square matrix and v, z are vectors satisfying:*[2]

- $v \neq 0$.
- *There is a scalar λ such that $Av = \lambda v$.*
- *There is a scalar $r \neq 0$ such that $Az = \lambda z + rv$.*

a) Show that v and z are both nonzero vectors in $Null((A - \lambda I)^2)$, where I is the identity matrix of the same size as A. The vector z is said to be a generalized eigenvector.

b) Show that $\{v, z\}$ are linearly independent.

c) Find such vectors $v, z \in \mathbb{R}^2$ and scalars $\lambda, r \in \mathbb{R}$ for $A = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$.

EXERCISE 3.28. *(Plotting sets) Define*

$$x_1 = (1, 2)$$
$$x_2 = (2, 1)$$
$$y = 5x_2 + (-4)x_1$$

a) Plot the points x_1, x_2, y in $\mathbb{R}^2$.

b) Draw a blob-like set $D \subseteq \mathbb{R}^2$ that contains x_1, x_2 but not y. Explain why D is not a subspace.

c) Explain why $Span\{x_1, x_2\} = \mathbb{R}^2$.

d) Explain why the line through points x_1, x_2 is obtained by $ax_1 + (1 - a)x_2$ for all $a \in \mathbb{R}$.

EXERCISE 3.29. *Define the following vectors $x, y \in \mathbb{R}^2$:*

$$x = \begin{bmatrix} 0 \\ 82 \end{bmatrix} \quad , \quad y = \begin{bmatrix} 14 \\ 3 \end{bmatrix}$$

a) Show that $\{x, y\}$ are linearly independent.

b) Express $(13, 3) \in \mathbb{R}^2$ as a linear combination of $\{x, y\}$.

c) Express $(-1, -1) \in \mathbb{R}^2$ as a linear combination of $\{x, y\}$.

[2] For the real field under the conditions of Exercise 3.27, it can be shown that matrix A is not *diagonalizable,* and the vector-based ODE $x'(t) = Ax(t)$ is solved by both $x_1(t) = ve^{\lambda t}$ and $x_2(t) = ze^{\lambda t} + ctve^{\lambda t}$ for some *particular* $c \in \mathbb{R}$ that is an easy exercise to find. See Chapter 14 for a treatment of ODEs of the type $x'(t) = Ax(t)$.

EXERCISE 3.30. *Define the following vectors $x, y \in \mathbb{R}^2$:*

$$x = \begin{bmatrix} 19 \\ 4 \end{bmatrix} \quad , \quad y = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$$

a) Show that $\{x, y\}$ are linearly independent.

b) Express $(13, 3) \in \mathbb{R}^2$ as a linear combination of $\{x, y\}$.

c) Express $(-1, -1) \in \mathbb{R}^2$ as a linear combination of $\{x, y\}$.

EXERCISE 3.31. *(Purely real basis) Consider the field of complex numbers $\mathbb{C}$. Recall that $i \in \mathbb{C}$ and $i^2 = -1$, and every $z \in \mathbb{C}$ has the form $z = a + bi$ for some $a, b \in \mathbb{R}$. Consider the vector space*

$$\mathbb{C}^n = \{(x_1, ..., x_n) : x_k \in \mathbb{C} \quad \forall k\}$$

We say that a subspace $S \subseteq \mathbb{C}^n$ has a purely real basis *if we can find a basis for S where all vectors of the basis are in $\mathbb{R}^n$.*

a) Find a purely real basis for $\mathbb{C}^2$.

b) Do all subspaces of $\mathbb{C}^2$ have a purely real basis? Prove true, or find a counter-example.

c) Do all subspaces of $\mathbb{C}^3$ have a purely real basis? Prove true, or find a counter-example.

EXERCISE 3.32. *(Complex conjugate vectors) Consider the complex field $\mathbb{C}$ and let $\mathbb{C}^2$ be the vector space of all 2-tuples of complex numbers. Define*

$$z = \begin{bmatrix} 4 + 3i \\ 2i \end{bmatrix} \quad , \quad \overline{z} = \begin{bmatrix} 4 - 3i \\ -2i \end{bmatrix}$$

Thus, $\overline{z}$ is the entrywise complex conjugate of z.

a) Prove that $\{z, \overline{z}\}$ are linearly independent.

b) Express the purely real vector $(8, 11) \in \mathbb{C}^2$ as a linear combination of $\{z, \overline{z}\}$.

c) Express the purely real vector $(-1, 3) \in \mathbb{C}^2$ as a linear combination of $\{z, \overline{z}\}$.

EXERCISE 3.33. *(Right-multiplication by an invertible matrix does not change $Col(A)$) Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Let G be an invertible $n \times n$ matrix (so the multiplication AG makes sense).*

a) Prove that $Col(A) = Col(AG)$.

b) What would change if G is not invertible?

EXERCISE 3.34. *(Linear independence in $\mathbb{R}^3$) Show the following vectors in $\mathbb{R}^3$ are linearly*

independent (or prove they are not):

$$\begin{bmatrix} 10 \\ 0 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 0 \\ 5 \end{bmatrix}$$

EXERCISE 3.35. *(Linear independence in $\mathbb{R}^3$) Show the following vectors in $\mathbb{R}^3$ are linearly independent (or prove they are not):*

$$\begin{bmatrix} 10 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \\ 5 \end{bmatrix}$$

EXERCISE 3.36. *(Linear independence in $\mathbb{R}^3$) Show the following vectors in $\mathbb{R}^3$ are linearly independent (or prove they are not):*

$$\begin{bmatrix} 10 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} -2 \\ 3 \\ 6 \end{bmatrix}$$

EXERCISE 3.37. *(Linear independence in $\mathbb{F}_2^3$) Let $\mathbb{F}_2 = \{0, 1\}$ be the binary field with mod 2 arithmetic. Show the following vectors in $\mathbb{F}_2^3$ are linearly independent (or prove they are not):*

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

EXERCISE 3.38. *(Linear independence in $\mathbb{F}_2^3$) Let $\mathbb{F}_2 = \{0, 1\}$ be the binary field with mod 2 arithmetic. Show the following vectors in $\mathbb{F}_2^3$ are linearly independent (or prove they are not):*

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

EXERCISE 3.39. *(Linear independence in $\mathbb{F}_3^3$) Let $\mathbb{F}_3 = \{0, 1, 2\}$ be the trinary field with mod 3*

arithmetic. Show the following vectors in $\mathbb{F}_3^3$ are linearly independent (or prove they are not):

$$\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

EXERCISE 3.40. *(Linear independence in $\mathbb{F}_3^3$) Let $\mathbb{F}_3 = \{0, 1, 2\}$ be the trinary field with mod 3 arithmetic. Show the following vectors in $\mathbb{F}_3^3$ are linearly independent (or prove they are not):*

$$\begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 0 \\ 1 \end{bmatrix}$$

EXERCISE 3.41. *(Linear combinations of linearly independent vectors) Suppose $\{x, y\}$ are linearly independent vectors in $\mathbb{F}^n$.*

a) Suppose $\mathbb{F}$ is a field for which $-1 \neq 1$. Show that $\{x + y, x - y\}$ are linearly independent.

b) Suppose $\mathbb{F} = \mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic. Show $\{x + y, x - y\}$ are linearly dependent.

EXERCISE 3.42. *(Linear combinations of linearly independent vectors) Suppose $\{x, y, z\}$ are linearly independent vectors in $\mathbb{F}^n$. Show that $\{x, x + y, x + y + z\}$ are linearly independent.*

EXERCISE 3.43. *(Linear combinations of linearly independent vectors) Suppose $\{x, y, z\}$ are linearly independent vectors in $\mathbb{R}^n$. Show that $\{2x, x + 8y, x + y - z\}$ are linearly independent.*

EXERCISE 3.44. *(Linear combinations of linearly independent vectors) Suppose $\{x, y, z\}$ are linearly independent in $\mathbb{R}^n$. Show that $\{x, 2x + y, 3x + y - z\}$ are linearly independent.*

EXERCISE 3.45. *(Linear combinations of linearly independent vectors) Suppose $\{x, y, z\}$ are linearly independent in $\mathbb{R}^n$. Show that $\{x + 8y, 3x + y + z, x - y - 2z\}$ are linearly independent.*

CHAPTER 4.

Gaussian elimination

Let $\mathbb{F}$ be a field. Let A be a $m \times n$ matrix with entries in $\mathbb{F}$. Fix $b \in \mathbb{F}^m$. We want to find all solutions $x \in \mathbb{F}^n$ to the equation

$$Ax = b \tag{4.1}$$

It is possible for there to be *no solutions*. It is possible for there to be *one unique solution*. It is also possible for there to be *more than one* solution. Let X^* denote the (possibly empty) set of all solutions to (4.1):

$$X^* = \{x \in \mathbb{F}^n : Ax = b\}$$

4.1. Particular solution and all solutions

Note that $X^* \subseteq \mathbb{F}^n$. To find set X^*, it is useful to study the closely related equation:

$$Ax = 0 \tag{4.2}$$

where the 0 on the right-hand-side is the all-zero vector in $\mathbb{F}^m$. Clearly $x = 0$ is *one particular solution* to (4.2), so the set of all solutions to (4.2) is nonempty. The set of all solutions to $Ax = 0$ already has a name, it is $Null(A)$ and is a subspace of $\mathbb{F}^n$.

Suppose we are fortunate enough to find a *particular solution* $x^* \in X^*$. So x^* is a particular vector in $\mathbb{F}^n$ that satisfies $Ax^* = b$. The next theorem shows we can find *all solutions* to $Ax = b$ just by knowing: (i) one particular solution x^*; (ii) the set $Null(A)$. We

show

$$X^* = x^* + Null(A)$$

where the right-hand-side is a sum of the particular solution x^* and the set $Null(A)$. What is the meaning of adding a vector to a set? We define the set $x^* + Null(A)$ according to the *Minkowski set sum* definition

$$x^* + Null(A) = \{x^* + z : z \in Null(A)\}$$

In other words, the set $x^* + Null(A)$ is formed by taking all vectors in $Null(A)$ and shifting them by the single vector x^*.

THEOREM 4.1. *Suppose x^* is a particular vector in $\mathbb{F}^n$ that satisfies $Ax^* = b$. Then a vector $x \in \mathbb{F}^n$ satisfies $Ax = b$ if and only if x has the form $x = x^* + z$ for some $z \in Null(A)$. Equivalently*

$$X^* = x^* + Null(A)$$

where $X^* = \{x \in \mathbb{F}^n : Ax = b\}$.

PROOF. First suppose x is a vector that satisfies $Ax = b$. We want to show $x = x^* + z$ for some $z \in Null(A)$. To this end, define

$$z = x - x^*$$

With this definition we clearly have $x = x^* + z$. It remains to show that $z \in Null(A)$:

$$\begin{aligned} Az &\overset{(a)}{=} A(x - x^*) \\ &\overset{(b)}{=} Ax - Ax^* \\ &\overset{(c)}{=} b - b \\ &= 0 \end{aligned}$$

where (a) substitutes the definition $z = x - x^*$; (b) distributes A across the difference of two vectors (recall (2.13)); (c) uses $Ax = b$ and $Ax^* = b$. Then $Az = 0$, so $z \in Null(A)$.

Conversely, suppose $x = x^* + z$ for some $z \in Null(A)$. We want to show $Ax = b$. We have

$$\begin{aligned} Ax &\overset{(a)}{=} A(x^* + z) \\ &= Ax^* + Az \\ &\overset{(b)}{=} b + 0 \end{aligned}$$

where (a) substitutes $x = x^* + z$; (b) uses $Ax^* = b$ and $Az = 0$. Thus, $Ax = b$. □

The above theorem guarantees that X^* is either the empty set, or has the structure of a shifted subspace. Such shifted subspaces are called *affine sets*. Affine sets are described more generally in Appendix C. See also Exercise 4.3.

4.1.1. Terminology

The equation $Ax = 0$ is called a *homogeneous equation* because its right-hand-side is 0. When $b \neq 0$, the equation $Ax = b$ is called an *inhomogeneous equation* because its right-hand-side is nonzero.

4.1.2. A line of solutions

Suppose x^* and w^* are two distinct vectors in the solution set $X^* = \{x \in \mathbb{F}^n : Ax = b\}$. That is, both x^* and w^* are vectors in $\mathbb{F}^n$, $x^* \neq w^*$, and

$$Ax^* = b \quad , \quad Aw^* = b$$

Claim: For all $\theta \in \mathbb{F}$ we have

$$\theta x^* + (1-\theta)w^* \in X^*$$

PROOF. Fix $\theta \in \mathbb{F}$. Then

$$\begin{aligned} A(\theta x^* + (1-\theta)w^*) &= \theta Ax^* + (1-\theta)Aw^* \\ &= \theta b + (1-\theta)b \\ &= b \end{aligned}$$

where the first equality distributes A according to the rule (2.13). □

In the special case when $\mathbb{F} = \mathbb{R}$, by sweeping through all infinitely many $\theta \in \mathbb{R}$, the vectors $\theta x^* + (1-\theta)w^*$ sweep out an infinite line in $\mathbb{R}^n$. Every point on this infinite line is itself a solution to $Ax = b$. Therefore, if there are at least two distinct solutions x^* and w^* to the equation $Ax = b$, then there must be *infinitely many* solutions. This motivates the following lemma.

LEMMA 4.1. *Suppose $\mathbb{F}$ is a field that has an infinite number of scalars. Let A be a $m \times n$ matrix with entries in $\mathbb{F}$. Fix $b \in \mathbb{F}^m$. The equation $Ax = b$ either has zero solutions, one solution, or infinitely many solutions.*

PROOF. If $Ax = b$ has zero solutions or one solution then we are done. Suppose it has at least two solutions. Let x^* and w^* be two distinct solutions, so $x^* \neq w^*$. We know that $\theta x^* + (1-\theta)w^*$ is also a solution for all $\theta \in \mathbb{F}$. We now show that each of the infinitely many scalars $\theta \in \mathbb{F}$ gives rise to a unique $\theta x^* + (1-\theta)w^*$ vector. Suppose not (we reach a

contradiction). Then there are two distinct scalars θ_1 and θ_2 for which

$$\theta_1 x^* + (1-\theta_1)w^* = \theta_2 x^* + (1-\theta_2)w^*$$

Rearranging this equation gives

$$(\theta_1 - \theta_2)(x^* - w^*) = 0 \tag{4.3}$$

Since the scalars θ_1 and θ_2 are distinct, we know $\theta_1 - \theta_2 \neq 0$ and the multiplicative inverse axiom ensures $(\theta_1 - \theta_2)^{-1}$ exists. Multiplying (4.3) by $(\theta_1 - \theta_2)^{-1}$ gives

$$(x^* - w^*) = 0$$

so $x^* = w^*$, which contradicts the assumption $x^* \neq w^*$. □

4.2. Back-substitution

Consider the following situation where solving $Ax = b$ is easy and leads to one and only one solution:

- A is square and upper triangular.
- All diagonal entries of A are nonzero.

4.2.1. Examples

1. Consider $\mathbb{F} = \mathbb{R}$ and

$$\underbrace{\begin{bmatrix} 1 & 3 & 7 \\ 0 & -1 & 4 \\ 0 & 0 & 2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 8 \\ 8 \\ 1 \end{bmatrix}}_{b}$$

   Write this as a system of 3 equations and 3 unknowns

$$\begin{aligned} x_1 + 3x_2 + 7x_3 &= 8 \\ -x_2 + 4x_3 &= 8 \\ 2x_3 &= 1 \end{aligned}$$

   Start from the bottom and solve up: From the third equation find x_3; Substitute the known numerical value for x_3 into the second equation to obtain x_2; Substitute the

known numerical values of (x_2, x_3) into the first equation to obtain x_1.

$$
\begin{aligned}
x_3 &= 1/2 \\
x_2 &= \frac{8 - 4x_3}{-1} = \frac{8 - 4(1/2)}{-1} = -6 \\
x_1 &= 8 - 3x_2 - 7x_3 = 8 - 3(-6) - 7(1/2) = 45/2
\end{aligned}
$$

Now test the answer to see if we made any mistakes:

$$
\begin{bmatrix} 1 & 3 & 7 \\ 0 & -1 & 4 \\ 0 & 0 & 2 \end{bmatrix}
\begin{bmatrix} 45/2 \\ -6 \\ 1/2 \end{bmatrix}
= \begin{bmatrix} 8 \\ 8 \\ 1 \end{bmatrix}
\quad \text{[correct]}
$$

This method is called *back-substitution*. [A similar solution method holds when the matrix A is lower triangular, we also call that back-substitution.] A computer can be programmed to do it very quickly. We show in the next subsection that it always leads to one and only one solution.

2. Consider $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic and

$$
\underbrace{\begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{A}
\underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}}_{x}
= \underbrace{\begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix}}_{b}
$$

Write this as a system of 4 equations and 4 unknowns

$$
\begin{aligned}
x_1 + x_2 + x_4 &= 1 \\
x_2 + x_4 &= 0 \\
x_3 + x_4 &= 1 \\
x_4 &= 1
\end{aligned}
$$

Start from the bottom and solve up:

$$
\begin{aligned}
x_4 &= 1 \\
x_3 &= 1 - x_4 = 1 + x_4 = 1 + 1 = 0 \\
x_2 &= 0 - x_4 = 0 + x_4 = 0 + 1 = 1
\end{aligned}
$$

$$x_1 = 1 - x_2 - x_4 = 1 + x_2 + x_4 = 1 + 1 + 1 = 1$$

Notice that $-1 = 1$ in mod 2 arithmetic, so $-x_4$ is the same as $+x_4$. Addition and subtraction are the same. Now test the answer to see if we made any mistakes:

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 1 \\ 1 \end{bmatrix} \quad \text{[correct]}$$

One way to check correctness with a computer is to do the usual matrix multiplication with integers, then take mod 2 of the output.

Listing 4.1. MATLAB for testing answer

```
A=[1 1 0 1;0 1 0 1;0 0 1 1;0 0 0 1]
mod(A*[1;1;0;1],2)
```

3. Consider $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic and

$$\underbrace{\begin{bmatrix} 2 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} 2 \\ 2 \\ 2 \\ 1 \end{bmatrix}}_{b}$$

Write this as a system of 4 equations and 4 unknowns

$$\begin{aligned} 2x_1 + x_2 + 2x_4 &= 2 \\ x_2 + x_4 &= 2 \\ x_3 + x_4 &= 2 \\ 2x_4 &= 1 \end{aligned}$$

For the mod 3 arithmetic it is useful to note that

$$-1 = 2 \quad , \quad -2 = 1 \quad , \quad 2^{-1} = 2$$

Start from the bottom and solve up:

$$
\begin{aligned}
x_4 &= 2^{-1} \cdot 1 = 2 \cdot 1 = 2 \\
x_3 &= 2 - x_4 = 2 - 2 = 0 \\
x_2 &= 2 - x_4 = 2 - 2 = 0 \\
x_1 &= 2^{-1}(2 - x_2 - 2x_4) = 2^{-1}(2 - 0 - (2)(2)) = 2(-2) = -4 = 2
\end{aligned}
$$

By properties of modular arithmetic, we can work with integers at intermediate steps, then at the end take everything mod 3. The final answer needs $x_i \in \{0, 1, 2\}$. Now test the answer to see if we made any mistakes:

$$
\begin{bmatrix} 2 & 1 & 0 & 2 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 2 \\ 0 \\ 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \\ 2 \\ 1 \end{bmatrix} \quad \text{[correct]}
$$

Listing 4.2. MATLAB for testing answer

```
A=[2 1 0 2;0 1 0 1;0 0 1 1;0 0 0 2]
mod(A*[2;0;0;2],3)
```

4.2.2. Why it works

The assumption that A is square, upper triangular, and has nonzero scalar entries on the diagonal ensures that $Ax = b$ has one and only one solution. Further, that solution is easily obtained by the back-substitution method. To see why, consider the general case $A = (a_{ij})$ with size $n \times n$, and $b = (b_i)$ with size n, with all entries in $\mathbb{F}$.

$$
\begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1,n-1} & a_{1n} \\ 0 & a_{22} & a_{23} & \cdots & a_{2,n-1} & a_{2n} \\ 0 & 0 & a_{33} & \cdots & a_{3,n-1} & a_{3n} \\ 0 & 0 & 0 & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & a_{n-1,n-1} & a_{n-1,n} \\ 0 & 0 & 0 & \cdots & 0 & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ \vdots \\ x_{n-1} \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \\ \vdots \\ b_{n-1} \\ b_n \end{bmatrix} \tag{4.4}
$$

Assume there is some vector $x = (x_1, \ldots, x_n)$ that solves the matrix equation (4.4). Since the scalar entries a_{ii} on the diagonal are nonzero, the multiplicative inverse axiom ensures

a_{ii}^{-1} exists for each $i \in \{1, \ldots, n\}$. The very last equation is

$$a_{nn}x_n = b_n \implies \boxed{x_n = a_{nn}^{-1}b_n} \tag{4.5}$$

so we always start back-substitution with an exactly computed value of x_n, being the only possible value that x_n can take.

Now use recursion (counting down from row n to row 1): Assume that when we operate on a row equation $i \in \{1, \ldots, n-1\}$, we have already computed numerical values of $x_{i+1}, x_{i+2}, \ldots, x_n$, and these are the only possible values they can take. The equation for row i is

$$a_{ii}x_i + \sum_{j=i+1}^{n} a_{ij}x_j = b_i$$

which means

$$x_i = \underbrace{a_{ii}^{-1}\left(b_i - \sum_{j=i+1}^{n} a_{ij}x_j\right)}_{\text{known}} \tag{4.6}$$

So x_i can be one and only one possible value, being the known value on the right-hand-side of the above equation. We now know the (only possible) numerical values of $x_i, x_{i+1}, \ldots, x_n$. Repeat the procedure until we finally compute x_1. Done.

Let us neglect any truncation or rounding errors that may occur in the numerical computations. Assume that the above back-substitution method performs exact computation with infinite precision. The above argument shows that in this situation:

- The back-substitution method always computes a specific vector $x = (x_1, \ldots, x_n)$.
- If a solution to the original problem $Ax = b$ exists, there is only one solution, and that is the one obtained by back-substitution.

A subtle question remains: Does a solution to $Ax = b$ exist? The second bullet point is a conditional statement that says *if* a solution exists, *then* it must be the one computed by back-substitution. How do we know that back-substitution is not giving us a *phantom output* x that does not really satisfy the $Ax = b$ equation? To appreciate the idea of a *phantom output*, the reader is encouraged to do Exercise 4.1 (see also the fourth bullet after Lemma 7.1 in Chapter 7).

The good news is that, in this situation of a square upper triangular matrix with nonzero entries on the diagonal, a solution to $Ax = b$ indeed always exists. One way to understand why is by reverse-engineering equation (4.5): While this equation is intended to *discover* the value of the scalar x_n that is already assumed to satisfy $a_{nn}x_n = b_n$, we can instead *define* $x_n = a_{nn}^{-1}b_n$ and then affirm that the equation $a_{nn}x_n = b_n$ is indeed true under this definition.

Similarly, equation (4.6), which *discovers* the value of x_i when we assume $(x_1, ..., x_n)$ solves (4.4), can be reverse-engineered to *define* x_i in terms of previously defined $x_{i+1}, ..., x_n$. Under this definition of x_i, the equation on row i in (4.4) indeed holds.

4.3. Overview of Gaussian elimination

We want a computational method that treats $Ax = b$ for any matrix A (not just square matrices). We want the complete solution set X^*. The general *Gaussian elimination* procedure does this in two phases:

1. Perform *invertible row operations* until A is in *echelon form*. Fortunately, any matrix A that is not an all-zero matrix can be put into echelon form.
2. Use back-substitution on echelon form matrices to get X^*. The method will either tell us that $X^* = \phi$ (where ϕ denotes the empty set), or will give a particular solution x^* together with a basis for $Null(A)$ (and of course $X^* = x^* + Null(A)$).

4.3.1. Dispensing with a trivial case

When describing Gaussian elimination, it is convenient to assume the matrix A is not all-zero. The all-zero case is trivial and we deal with that now: Define $0_{m\times n}$ as the all-zero matrix of size $m \times n$. Suppose $A = 0_{m\times n}$. Then

$$0_{m\times n}x = 0 \quad \forall x \in \mathbb{F}^n$$

which means the set of all solutions to $Ax = b$ is

$$X^* = \begin{cases} \phi & \text{if } b \neq 0 \\ \mathbb{F}^n & \text{if } b = 0 \end{cases}$$

That is, if $b \neq 0$ then there are no solutions to $0_{m\times n}x = b$. If $b = 0$ then all vectors in the vector space $\mathbb{F}^n$ are solutions.

4.3.2. The theory of Gaussian elimination

This theorem provides the theoretical foundation.

THEOREM 4.2. *Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Fix $b \in \mathbb{F}^m$.*

a) Let G be any $k \times m$ matrix with entries in $\mathbb{F}$. Then

$$\{x \in \mathbb{F}^n : Ax = b\} \subseteq \{x \in \mathbb{F}^n : GAx = Gb\}$$

Thus, left-multiplying both sides of the equation $Ax = b$ by a matrix G either maintains the same solution set, or expands it.

b) If G is a $m \times m$ invertible matrix (with entries in $\mathbb{F}$) then

$$\{x \in \mathbb{F}^n : Ax = b\} = \{x \in \mathbb{F}^n : GAx = Gb\}$$

Thus, left-multiplying both sides of the equation $Ax = b$ by an invertible matrix G does not change the solution set.

PROOF. To prove (a), suppose x is a vector that satisfies

$$Ax = b$$

Multiplying both sides of the above equation by G gives

$$GAx = Gb$$

and this completes the proof of part (a).

To prove (b), we already know $\{x \in \mathbb{F}^n : Ax = b\} \subseteq \{x \in \mathbb{F}^n : GAx = Gb\}$. It suffices to prove the reverse inclusion. Let x be a vector that satisfies

$$GAx = Gb$$

Multiplying both sides by G^{-1} gives

$$\underbrace{G^{-1}G}_{I}Ax = \underbrace{G^{-1}G}_{I}b$$

Thus

$$IAx = Ib$$

But $IA = A$ and $Ib = b$ and so $Ax = b$. □

Gaussian elimination seeks to solve the equation $Ax = b$. It first performs *row reductions* that are equivalent to left-multiplying the equation by a sequence of invertible matrices $G_1, G_2, ..., G_k$, where k is some positive integer number of steps:

- First step: $G_1Ax = G_1b$. [The solution set does not change]
- Second step: $G_2G_1Ax = G_2G_1b$. [The solution set does not change]
- Third step: $G_3G_2G_1Ax = G_3G_2G_1b$. [The solution set does not change]

and so on. At the end we have:

$$\underbrace{G_k G_{k-1} \cdots G_1}_{G} Ax = \underbrace{G_k G_{k-1} \cdots G_1}_{G} b$$

where k is the total number of steps. The product of the invertible G_i matrices used at each step is again an invertible matrix, call it G:

$$G = G_k G_{k-1} \cdots G_1$$

Thus, we know

$$X^* = \{x : Ax = b\} = \{x : GAx = Gb\}$$

The desired solution set X^* is preserved by this sequence of left-multiplications. What invertible G_i matrices should be used at each step? The Gaussian elimination technique uses G_i matrices that are designed to turn *any* $m \times n$ matrix A that is not an all-zero matrix into a $m \times n$ matrix R with a particular structure called *echelon form*:

$$GA = G_k G_{k-1} \cdots G_1 A = R$$

Then

$$X^* = \{x : Ax = b\} = \{x : GAx = Gb\} = \{x : Rx = Gb\}$$

and the echelon form of R makes solving $Rx = c$ easy for any vector c (including the vector $c = Gb$). The next section defines echelon form and shows why solving $Rx = c$ is easy.

4.4. Echelon form

We describe a matrix in echelon form. The matrix is assumed to have entries in some field $\mathbb{F}$, and not all entries are zero.

4.4.1. Basic echelon form

A $m \times n$ matrix R that is not an all-zero matrix is in *echelon form*, also called *basic echelon form*, if

- There is an integer $r \in \{1, \ldots, m\}$, called the *rank* of R, such that the first r rows of R are not all-zero rows, while the remaining $m - r$ rows (if any) are all-zero rows.
- The first nonzero entry on each row $i \in \{1, \ldots, r\}$ is called a *pivot*.
- For each $i \in \{2, \ldots, r\}$, the pivot on row i is in a column that is at least one column to the right of the pivot on row $i - 1$.

If R is a square matrix in echelon form, then it must be upper triangular. For general (possibly nonsquare) matrices R, columns without pivots are called *free variable columns*. When solving $Rx = c$ for a vector $x = (x_1, \ldots, x_n)$, an entry x_i is called a *free variable* if i is the index number of a free variable column. If R has rank r, then there are r pivots and $n - r$ free variables. We use notation $rank(R)$ to denote the rank of matrix R:

$$rank(R) = r = \text{number of pivots}$$
$$n - r = \text{number of free variables}$$

Echelon form makes solving $Rx = c$ easy via back-substitution. Examples:

1. Using field $\mathbb{R}$:

$$R = \begin{bmatrix} 0 & \boxed{2} & 1 & 3 & 1 & 3 \\ 0 & 0 & 0 & \boxed{-8} & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4.7}$$

   There are two pivots (they are boxed). We have $r = rank(R) = 2$ and the number of free variables is $6 - 2 = 4$. Free variables occur in columns 1, 3, 5, 6. We say the free variables are x_1, x_3, x_5, x_6. Now let us solve $Rx = c$ for a general $c \in \mathbb{R}^4$:

$$R = \begin{bmatrix} 0 & \boxed{2} & 1 & 3 & 1 & 3 \\ 0 & 0 & 0 & \boxed{-8} & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$$

   This matrix equation is equivalent to the following system of 4 equations:

$$2x_2 + x_3 + 3x_4 + x_5 + 3x_6 = c_1 \tag{4.8}$$
$$-8x_4 + x_6 = c_2 \tag{4.9}$$
$$0 = c_3 \tag{4.10}$$
$$0 = c_4 \tag{4.11}$$

If $c_3 \neq 0$ or $c_4 \neq 0$, it is clear that there is *no solution,* so $X^* = \phi$. We shall show that

$$(X^* = \phi) \iff (c_3 \neq 0 \text{ or } c_4 \neq 0)$$

Assume $c_3 = c_4 = 0$. Starting with equation (4.9), we solve for x_4 in terms of the free variables x_i for $i > 4$:

$$x_4 = \frac{c_2 - x_6}{-8} = (-1/8)c_2 + (1/8)x_6$$

Then by back-substitution into (4.8)

$$\begin{aligned} x_2 &= \frac{c_1 - x_3 - 3x_4 - x_5 - 3x_6}{2} \\ &= \frac{c_1 - x_3 - 3\left((-1/8)c_2 + (1/8)x_6\right) - x_5 - 3x_6}{2} \\ &= (1/2)c_1 + (3/16)c_2 + (-1/2)x_3 + (-1/2)x_5 + (-27/16)x_6 \end{aligned}$$

So a general solution can be written in terms of the free variables x_1, x_3, x_5, x_6:

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{bmatrix} = \begin{bmatrix} x_1 \\ (1/2)c_1 + (3/16)c_2 + (-1/2)x_3 + (-1/2)x_5 + (-27/16)x_6 \\ x_3 \\ (-1/8)c_2 + (1/8)x_6 \\ x_5 \\ x_6 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} 0 \\ (1/2)c_1 + (3/16)c_2 \\ 0 \\ (-1/8)c_2 \\ 0 \\ 0 \end{bmatrix}}_{x^*} + x_1 \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{z_1} + x_3 \underbrace{\begin{bmatrix} 0 \\ -1/2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{z_2} + x_5 \underbrace{\begin{bmatrix} 0 \\ -1/2 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}}_{z_3} + x_6 \underbrace{\begin{bmatrix} 0 \\ -27/16 \\ 0 \\ 1/8 \\ 0 \\ 1 \end{bmatrix}}_{z_4}$$

The above equation expresses every possible solution x to the equation $Rx = c$ in terms of the 4 free variables x_1, x_3, x_5, x_6. These variables can be chosen as *any desired real numbers* and the resulting vector x still solves $Rx = c$. The simplest solution is obtained by choosing all free variables to be 0, so $x_1 = x_3 = x_5 = x_6 = 0$. Then the above expression for x reduces to $x = x^*$. So x^* is a particular solution to $Rx = c$. Also, the only way for $x^* + x_1 z_1 + x_3 z_2 + x_5 z_3 + x_6 z_4$ to *also* be a solution, for all possible

$x_1, x_3, x_5, x_6 \in \mathbb{R}$, is if

$$Rz_i = 0 \quad \forall i \in \{1, 2, 3, 4\}$$

This is formally proven in Exercise 4.12 (Hint: To prove $Rz_1 = 0$, just choose $x_1 = 1$ and $x_3 = x_5 = x_6 = 0$.) Thus, $z_i \in Null(R)$ for all $i \in \{1, 2, 3, 4\}$.
Overall, we have obtained a *particular solution* x^* and a *basis for the null space* $\{z_1, z_2, z_3, z_4\}$ (these can be shown to be linearly independent, see Exercise 4.12). So $Null(R) = Span\{z_1, z_2, z_3, z_4\}$ and

$$\boxed{X^* = \{x : Rx = c\} = \begin{cases} x^* + Span\{z_1, z_2, z_3, z_4\} & \text{if } c_3 = c_4 = 0 \\ \phi & \text{else} \end{cases}}$$

In general, this procedure always leads to linearly independent vectors that are a basis for the null space. The number of vectors in a basis for the null space is always the number of free variables $n - r$.
To ensure there are no mistakes, we need to test that $Rz_i = 0$ for all $i \in \{1, 2, 3, 4\}$:

$$Rz_1 = Rz_2 = Rz_3 = Rz_4 = 0 \quad \text{[correct]}$$

We also need to test the particular solution to verify $Rx^* = (c_1, c_2, 0, 0)$.

2. Using $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic:

$$R = \begin{bmatrix} \boxed{1} & 1 & 1 & 0 & 1 \\ 0 & 0 & \boxed{1} & 1 & 0 \\ 0 & 0 & 0 & \boxed{1} & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

There are 3 pivots. The rank is $r = 3$. The number of free variables is $5 - 3 = 2$:

$$\text{Free variables:} \quad x_2, x_5$$

Let us solve $Rx = c$ for a general $c \in \{0, 1\}^4$:

$$R = \begin{bmatrix} \boxed{1} & 1 & 1 & 0 & 1 \\ 0 & 0 & \boxed{1} & 1 & 0 \\ 0 & 0 & 0 & \boxed{1} & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$$

This is equivalent to the system of equations

$$x_1 + x_2 + x_3 + x_5 = c_1 \tag{4.12}$$

$$x_3 + x_4 = c_2 \tag{4.13}$$

$$x_4 + x_5 = c_3 \tag{4.14}$$

$$0 = c_4 \tag{4.15}$$

Clearly there is no solution if $c_4 = 1$. It turns out that there is no solution if and only if $c_4 = 1$, as we now show. Assuming $c_4 = 0$, we start with (4.14) to obtain x_4 in terms of the free variable x_5 (using mod 2 arithmetic where $-1 = 1$):

$$x_4 = c_3 - x_5 = c_3 + x_5$$

We then perform back-substitution, starting with equation (4.13), to solve for x_3 and x_1 in terms of the free variables x_2, x_5:

$$\begin{aligned}
x_3 &= c_2 - x_4 = c_2 + x_4 = c_2 + (c_3 + x_5) \\
x_1 &= c_1 - x_2 - x_3 - x_5 \\
&= c_1 + x_2 + x_3 + x_5 \\
&= c_1 + x_2 + (c_2 + (c_3 + x_5)) + x_5 \\
&= c_1 + c_2 + c_3 + x_2
\end{aligned}$$

which uses $x_5 + x_5 = (1 + 1)x_5 = 0$, since $2 = 0$ in the binary field. Thus

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} c_1 + c_2 + c_3 + x_2 \\ x_2 \\ c_2 + c_3 + x_5 \\ c_3 + x_5 \\ x_5 \end{bmatrix}$$

$$= \underbrace{\begin{bmatrix} c_1 + c_2 + c_3 \\ 0 \\ c_2 + c_3 \\ c_3 \\ 0 \end{bmatrix}}_{x^*} + x_2 \underbrace{\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{z_1} + x_5 \underbrace{\begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \\ 1 \end{bmatrix}}_{z_2}$$

We have obtained a particular solution x^* and a basis $\{z_1, z_2\}$ for $Null(R)$. Thus

$$\boxed{X^* = \{x : Rx = c\} = \begin{cases} x^* + Span\{z_1, z_2\} & \text{if } c_4 = 0 \\ \phi & \text{if } c_4 = 1 \end{cases}}$$

We should test this solution to see if we made any mistakes:

$$Rz_1 = Rz_2 = 0 \quad [\text{correct}]$$

$$Rx^* = (c_1, c_2, c_3, 0) \quad [\text{correct}]$$

$$\text{dimension of } Null(R) = \text{number of free variables} \quad [\text{correct}]$$

4.4.2. Row reduced echelon form

We say a $m \times n$ matrix R is in *row reduced echelon form* if it is in basic echelon form and

- All pivot entries are 1.
- All columns with pivots are zeros in every row, except for the pivot row.

All nonzero $m \times n$ matrices can be turned into a row reduced echelon form matrix by a sequence of elementary row operations, each operation corresponding to left-multiplication by an invertible matrix. (The procedure is described in the next section.) Once we reach basic echelon form we can stop and perform back-substitution. Or, we can keep going to row reduced echelon form, which takes more steps but simplifies the back-substitution.

An example of row reduced echelon form is (using field $\mathbb{R}$):

$$R = \begin{bmatrix} \boxed{1} & 2 & 0 & 0 & -8 \\ 0 & 0 & \boxed{1} & 0 & 3 \\ 0 & 0 & 0 & \boxed{1} & 5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

The pivots are boxed. We have $rank(R) = r = 3$. There are $5 - 3 = 2$ free variables. The free variables are

$$\text{free variables:} \quad x_2 \text{ and } x_5$$

Consider solving $Rx = c$:

$$\underbrace{\begin{bmatrix} 1 & 2 & 0 & 0 & -8 \\ 0 & 0 & 1 & 0 & 3 \\ 0 & 0 & 0 & 1 & 5 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}}_{R} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \\ c_4 \end{bmatrix}$$

This is equivalent to the system of equations

$$\begin{aligned} x_1 + 2x_2 - 8x_5 &= c_1 \\ x_3 + 3x_5 &= c_2 \\ x_4 + 5x_5 &= c_3 \\ 0 &= c_4 \end{aligned}$$

The solution set $X^* = \phi$ if and only if $c_4 \neq 0$. Assuming $c_4 = 0$, we obtain

$$\begin{aligned} x_1 &= c_1 - 2x_2 + 8x_5 \\ x_2 &= x_2 \\ x_3 &= c_2 - 3x_5 \\ x_4 &= c_3 - 5x_5 \\ x_5 &= x_5 \end{aligned}$$

So the set of all solutions is

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \underbrace{\begin{bmatrix} c_1 \\ 0 \\ c_2 \\ c_3 \\ 0 \end{bmatrix}}_{x^*} + x_2 \underbrace{\begin{bmatrix} -2 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{z_1} + x_5 \underbrace{\begin{bmatrix} 8 \\ 0 \\ -3 \\ -5 \\ 1 \end{bmatrix}}_{z_2}$$

for all possible choices of the free variables $x_2, x_5 \in \mathbb{R}$. A particular solution is x^*. This has a simple form where the particular solution x^* is a vector that has either 0 entries, or c_i

entries for some c_i of the c vector. A basis for $Null(R)$ is $\{z_1, z_2\}$. We can test this

$$Rz_1 = Rz_2 = 0 \quad \text{[correct]}$$

$$Rx^* = (c_1, c_2, c_3, 0) \quad \text{[correct]}$$

4.5. Elementary row operations

Consider the equation $Ax = b$, where A is a $m \times n$ matrix with entries in some field $\mathbb{F}$ and $b \in \mathbb{F}^m$. Assume A is not an all-zero matrix. Define the solution set

$$X^* = \{x : Ax = b\}$$

We use three types of row operations, each one related to left-multiplication by a certain type of invertible matrix G. All examples are illustrated for the special case equation

$$\underbrace{\begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix}}_{x} = \underbrace{\begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix}}_{b} \tag{4.16}$$

It shall be convenient to write matrix equations of the type (4.16) in the following *augmented matrix* form:

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right] \tag{4.17}$$

The entire matrix equation (4.16) is understood to be represented by the above augmented matrix (we have removed the vector $x = (x_1, x_2, x_3, x_4)$; we have also removed the equality symbol "="). The dividing line in (4.17) makes it clear which parts of the augmented matrix correspond to the matrix A and which correspond to b. This is a mere notational convenience: It is easy to go back and forth between the matrix equation (4.16) and its augmented matrix form (4.17).

The matrix equation (4.16), and its augmented form (4.17), are equivalent to the following system of 3 equations and 4 unknowns:

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = b_1 \tag{4.18}$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = b_2 \tag{4.19}$$

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 = b_3 \tag{4.20}$$

The solution set X^* for the equation (4.16) is the same as for the system (4.18)-(4.20).

4.5.1. Type 1 row operation: permuting a row

The solution set to the system of equations (4.18)-(4.20) is unchanged if the order of equations 1 and 3 is swapped:

$$a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 = b_3 \tag{4.21}$$

$$a_{21}x_1 + a_{22}x_2 + a_{23}x_3 + a_{24}x_4 = b_2 \tag{4.22}$$

$$a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 = b_1 \tag{4.23}$$

Swapping rows is an *invertible row operation* because it does not lose any information; we can always swap those same rows back to obtain the original system. For example, system (4.18)-(4.20) can be recovered from system (4.21)-(4.23) by swapping the ordering of equations (4.21) and (4.23).

A general $Ax = b$ equation is equivalent to a system of m equations and n unknowns. The solution to the system of equations is unchanged if the order of two equations is swapped. Swapping two equations corresponds to swapping two rows in the augmented matrix. For example, swapping rows 1 and 3 in (4.17) gives

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right] \rightarrow \left[\begin{array}{cccc|c} a_{31} & a_{32} & a_{33} & a_{34} & b_3 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{11} & a_{12} & a_{13} & a_{14} & b_1 \end{array}\right]$$

This procedure of swapping rows in the augmented matrix is called a *type 1 invertible row operation*, also called a *type 1 elementary row operation*. The same result can be obtained by left-multiplying the original augmented matrix by a $m \times m$ *permutation matrix*, which is a 3×3 permutation matrix P for this example:

$$\underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \end{bmatrix}}_{P} \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right] = \left[\begin{array}{cccc|c} a_{31} & a_{32} & a_{33} & a_{34} & b_3 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{11} & a_{12} & a_{13} & a_{14} & b_1 \end{array}\right]$$

In general, the permutation matrix needed to swap rows i and j in a $m \times n$ matrix is the $m \times m$ matrix P formed by swapping rows i and j of the $m \times m$ identity matrix I. Left-

multiplying any matrix with m rows by this matrix P swaps rows i and j of that matrix. Since P itself is a matrix with m rows, left-multiplying P with itself swaps those rows back to give the identity matrix:

$$P^2 = PP = I$$

It follows that matrix P is invertible, and its inverse is P:

$$P^{-1} = P$$

Summary: Swapping two rows i and j is the first type of elementary row operation. It does not change the solution set. Swapping rows can be accomplished in the augmented matrix by left-multiplying by an invertible permutation matrix P. The matrix P is formed by swapping rows i and j of the identity matrix I.

4.5.2. Type 2 row operation: Multiplying a row by a scalar $\theta \neq 0$

The second type of elementary row operation chooses any desired scalar $\theta \neq 0$ and multiplies any desired row of the augmented matrix by θ. For example, we can multiply row 2 by θ in our example system:

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right] \rightarrow \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ \theta a_{21} & \theta a_{22} & \theta a_{23} & \theta a_{24} & \theta b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right] \tag{4.24}$$

This is an invertible row operation because it does not lose any information: Since the scalar θ is nonzero, the multiplicative inverse axiom ensures θ^{-1} exists. Thus, we can undo what we have done simply by multiplying row 2 of the matrix on the right-hand-side of (4.24) by θ^{-1}. It is important that $\theta \neq 0$, else the operation in (4.24) would wipe out all information on row 2 of the original augmented matrix! Multiplying row 2 by the scalar 0 would be a *noninvertible* row operation, and is not allowed.

Assuming $\theta \neq 0$, the solution set for the new augmented matrix has not changed. Indeed, the new augmented matrix (on the right-hand-side of (4.24)) can be translated into its corresponding system of equations

$$\begin{aligned} a_{11}x_1 + a_{12}x_2 + a_{13}x_3 + a_{14}x_4 &= b_1 \\ \theta a_{21}x_1 + \theta a_{22}x_2 + \theta a_{23}x_3 + \theta a_{24}x_4 &= \theta b_2 \\ a_{31}x_1 + a_{32}x_2 + a_{33}x_3 + a_{34}x_4 &= b_3 \end{aligned}$$

The only difference between this new system of equations and (4.18)-(4.20) is that the second equation has been multiplied by the nonzero scalar θ. It is clear that the solution

set is not changed.

The elementary row operation of multiplying a given row $i \in \{1, ..., m\}$ of the augmented matrix by a nonzero scalar θ can be accomplished by left-multiplying the augmented matrix by a square matrix G that is formed from the $m \times m$ identity matrix I by multiplying row i of I by θ:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G} \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right] = \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ \theta a_{21} & \theta a_{22} & \theta a_{23} & \theta a_{24} & \theta b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right]$$

Since $\theta \neq 0$, the G matrix is invertible: Its inverse is formed by taking the identity matrix I and multiplying that same row by θ^{-1}:

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \theta^{-1} & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G^{-1}} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{I}$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \theta^{-1} & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G^{-1}} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & \theta & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{I}$$

Summary: Multiplying a row by a nonzero scalar θ is the second type of elementary row operation. It does not change the solution set. This operation can be accomplished on the augmented matrix by left-multiplying by an invertible matrix G. The matrix G is formed by multiplying that same row of the identity matrix I by θ.

4.5.3. Type 3 row operation: θ(row i) + (row j) $\to$ (row j)

The third type of elementary row operation takes any desired scalar θ and any two distinct rows i and j of the augmented matrix. Then, it replaces row j with

$$\theta(\text{row } i) + (\text{row } j)$$

For example, taking row $i = 1$ and $j = 2$ gives

$$\left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right]$$

$$\to \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ (\theta a_{11} + a_{21}) & (\theta a_{12} + a_{22}) & (\theta a_{13} + a_{23}) & (\theta a_{14} + a_{24}) & (\theta b_1 + b_2) \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right] \tag{4.25}$$

This is an invertible row operation because it does not lose any information: If we are given the augmented matrix on the right-hand-side of (4.25), we can recover the original augmented matrix by multiplying row 1 of the new augmented matrix by $-\theta$ and adding it into row 2 of the new augmented matrix. If $\theta = 0$ this is still an invertible row operation, but in that case the new row is the same as the old.

This operation can be accomplished by left-multiplying the original augmented matrix by an invertible matrix G. That G is formed by performing the same desired row operation on I (multiply row i of I by θ and add into row j of I).

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \theta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G} \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ a_{21} & a_{22} & a_{23} & a_{24} & b_2 \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right]$$

$$= \left[\begin{array}{cccc|c} a_{11} & a_{12} & a_{13} & a_{14} & b_1 \\ (\theta a_{11} + a_{21}) & (\theta a_{12} + a_{22}) & (\theta a_{13} + a_{23}) & (\theta a_{14} + a_{24}) & (\theta b_1 + b_2) \\ a_{31} & a_{32} & a_{33} & a_{34} & b_3 \end{array}\right]$$

Such a matrix G is invertible: Its inverse G^{-1} is the same as G with the exception that the entry with θ is replaced by $-\theta$.

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \theta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -\theta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G^{-1}} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ -\theta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G^{-1}}\underbrace{\begin{bmatrix} 1 & 0 & 0 \\ \theta & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}_{G} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

Summary: A type 3 elementary row operation multiplies some row i by a scalar θ and adds the result into some other row j. This is an invertible row operation. It does not change the solution set. This operation can be accomplished on the augmented matrix by left-multiplying by an invertible matrix G. The matrix G is formed by multiplying row i of the identity matrix by θ and adding the result into row j of the identity matrix.

4.6. How an A becomes an R

Start with a matrix A that is not all-zero. Then

- Find a row that has a left-most nonzero entry (breaking ties arbitrarily). Row swap that into row 1. Its nonzero entry becomes the first pivot. Let p_1 be the entry value of that pivot (we know $p_1 \neq 0$).
- Using type 3 operations, turn all entries below the first pivot into zero. To do this, if there is another row with an entry $\alpha \neq 0$ in the same column of that first pivot, multiply row 1 by $-p_1^{-1}\alpha$ and add it into that other row (p_1^{-1} exists since $p_1 \neq 0$). This turns α into 0.
- Once the column for pivot 1 is taken care of, we are done with row 1. Repeat the procedure on the remaining nonzero rows of the matrix. In principle, we can strip away all columns on or to the left of that first pivot column, also strip away the entire first row, then treat the remaining (smaller) matrix according to the first step above. Keep going until we reach echelon form.

4.7. Examples

1. Consider field $\mathbb{R}$.

$$\begin{bmatrix} 0 & 3 & -3 & 2 \\ 1 & 1 & 1 & 1 \\ 2 & 9 & 0 & 2 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

Form the augmented matrix and perform elementary row operations. The first step is to swap rows 1 and 2:

$$\left[\begin{array}{cccc|c} 0 & 3 & -3 & 2 & 1 \\ 1 & 1 & 1 & 1 & 2 \\ 2 & 9 & 0 & 2 & 3 \end{array}\right] \to \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 2 \\ 0 & 3 & -3 & 2 & 1 \\ 2 & 9 & 0 & 2 & 3 \end{array}\right] \quad \text{[now (-2)(row 1)+(row 3)→ (row3)]}$$

$$\to \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 2 \\ 0 & 3 & -3 & 2 & 1 \\ 0 & 7 & -2 & 0 & -1 \end{array}\right] \quad \text{[now multiply row 2 by 1/3]}$$

$$\to \left[\begin{array}{cccc|c} 1 & 1 & 1 & 1 & 2 \\ 0 & 1 & -1 & 2/3 & 1/3 \\ 0 & 7 & -2 & 0 & -1 \end{array}\right] \quad \text{[now (-7)(row2) + (row 3) → (row 3)]}$$

$$\to \left[\begin{array}{cccc|c} \boxed{1} & 1 & 1 & 1 & 2 \\ 0 & \boxed{1} & -1 & 2/3 & 1/3 \\ 0 & 0 & \boxed{5} & -14/3 & -10/3 \end{array}\right]$$

This is in basic echelon form so we can stop here. There are $r = 3$ pivots. There is one free variable: x_4. The final augmented matrix is equivalent to

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 2 \\ x_2 - x_3 + (2/3)x_4 &= 1/3 \\ 5x_3 + (-14/3)x_4 &= -10/3 \end{aligned}$$

Therefore

$$\begin{aligned} x_3 &= -10/15 + (14/15)x_4 \\ x_2 &= 1/3 + x_3 - (2/3)x_4 \\ &= 1/3 + (-10/15 + (14/15)x_4) - (2/3)x_4 \\ &= -1/3 + (4/15)x_4 \\ x_1 &= 2 - x_2 - x_3 - x_4 \\ &= 2 - (-1/3 + (4/15)x_4) - (-10/15 + (14/15)x_4) - x_4 \\ &= 3 + (-11/5)x_4 \end{aligned}$$

So

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \underbrace{\begin{bmatrix} 3 \\ -1/3 \\ -10/15 \\ 0 \end{bmatrix}}_{x^*} + x_4 \underbrace{\begin{bmatrix} -11/5 \\ 4/15 \\ 14/15 \\ 1 \end{bmatrix}}_{z}$$

So

$$\boxed{X^* = x^* + Span\{z\}}$$

We can verify that, indeed, $Az = 0$ and $Ax^* = (1, 2, 3)$.

2. Consider field $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic.

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

Note that $1 + 1 = 0$ and $-1 = 1$. The first step is to swap rows 1 and 2:

$$\left[\begin{array}{ccccc|c} 0 & 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccccc|c} 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{array}\right]$$ [now add row 2 into row 3]

$$\to \left[\begin{array}{ccccc|c} 1 & 1 & 0 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right]$$ [this is basic echelon form, but let us continue] [now add row 2 into row 1]

$$\to \left[\begin{array}{ccccc|c} 1 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{array}\right]$$ [now add row 3 into row 2]

$$\to \left[\begin{array}{ccccc|c} \boxed{1} & 0 & 1 & 1 & 0 & 0 \\ 0 & \boxed{1} & 1 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & \boxed{1} & 0 \end{array}\right]$$ [this is row reduced echelon form]

The rank is $r = 3$. There are 2 free variables: x_3, x_4. We have

$$x_1 + x_3 + x_4 = 0 \implies x_1 = x_3 + x_4$$
$$x_2 + x_3 = 1 \implies x_2 = 1 + x_3$$
$$x_5 = 0 \implies x_5 = 0$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}}_{x^*} + x_3 \underbrace{\begin{bmatrix} 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}}_{z_1} + x_4 \underbrace{\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}}_{z_2}$$

So x^* is a particular solution and $Null(A) = Span\{z_1, z_2\}$.

$$\boxed{X^* = x^* + Span\{z_1, z_2\}}$$

Verify:

$$Az_1 = Az_2 = 0 \quad [\text{correct}]$$

$$Ax^* = (1, 1, 1) \quad [\text{correct}]$$

Policy of thumb: To convert to row reduced echelon form, it is typically most efficient to first convert to echelon form (changing only rows underneath a given row, starting from the top); then work towards row reduced echelon form by clearing the nonzero entries above the pivots *starting from the bottom row and working up*. Once echelon form is reached, the remaining steps that start from the bottom are easier because those bottom rows have more zeros. Also, creating zeros above them makes the future upward-working steps easier. This is often how computer implementations of Gaussian elimination work. The procedure illustrated above *breaks this policy* in its "but let us continue" step by adding row 2 into row 1, rather than starting from the bottom and adding row 3 into row 2. In this case it works well since it accidentally creates *two* needed zeros on row 1, instead of one. Despite this savings, perhaps it would have been better pedagogical form if the example were carried out in a way that did not break the policy we are now suggesting. However, it illustrates there are multiple valid ways to perform row reductions, some ways having more steps than others.

3. Consider $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic.

$$\begin{bmatrix} 2 & 2 & 0 & 2 \\ 0 & 1 & 2 & 0 \\ 1 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \\ 2 \end{bmatrix}$$

Note that

$$2^{-1} = 2, \quad -2 = 1, \quad -1 = 2, \quad 2 + 1 = 0, \quad -4 = 2$$

The first step is to add row 1 into row 3:

$$\left[\begin{array}{cccc|c} 2 & 2 & 0 & 2 & 0 \\ 0 & 1 & 2 & 0 & 1 \\ 1 & 0 & 1 & 0 & 2 \end{array}\right] \to \left[\begin{array}{cccc|c} 2 & 2 & 0 & 2 & 0 \\ 0 & 1 & 2 & 0 & 1 \\ 0 & 2 & 1 & 2 & 2 \end{array}\right] \quad \text{[now add row 2 into row 3]}$$

$$\to \left[\begin{array}{cccc|c} 2 & 2 & 0 & 2 & 0 \\ 0 & 1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 2 & 0 \end{array}\right] \quad \begin{array}{l} \text{[this is basic echelon form, but let us continue]} \\ \text{[now 2(row3) +(row1)→ (row 1)]} \end{array}$$

$$\to \left[\begin{array}{cccc|c} 2 & 2 & 0 & 0 & 0 \\ 0 & 1 & 2 & 0 & 1 \\ 0 & 0 & 0 & 2 & 0 \end{array}\right] \quad \text{[now add row 2 into row 1]}$$

$$\to \left[\begin{array}{cccc|c} \boxed{2} & 0 & 2 & 0 & 1 \\ 0 & \boxed{1} & 2 & 0 & 1 \\ 0 & 0 & 0 & \boxed{2} & 0 \end{array}\right]$$

The rank is $r = 3$. There is 1 free variable: x_3.

$$2x_1 + 2x_3 = 1 \implies x_1 = 2^{-1}(1 - 2x_3) = 2(1 - 2x_3) = 2 - 4x_3 = 2 + 2x_3$$
$$x_2 + 2x_3 = 1 \implies x_2 = 1 - 2x_3 = 1 + x_3$$
$$2x_4 = 0 \implies x_4 = 0$$

$$\begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \underbrace{\begin{bmatrix} 2 \\ 1 \\ 0 \\ 0 \end{bmatrix}}_{x^*} + x_3 \underbrace{\begin{bmatrix} 2 \\ 1 \\ 1 \\ 0 \end{bmatrix}}_{z}$$

So x^* is a particular solution, $Null(A) = Span\{z\}$, and

$$X^* = x^* + Span\{z\}$$

We can verify the solution:

$$Az = 0 \quad \text{[correct]}$$

$$Ax^* = (0, 1, 2) \quad \text{[correct]}$$

4.8. Computing A^{-1}

The following *Gauss-Jordan* method computes the inverse of a square and invertible matrix A by Gaussian elimination. It can be shown that if A and B are square matrices that satisfy either $AB = I$ or $BA = I$, then both A and B are invertible (and they are inverses of each other). This is formally proven in the next chapter. For now, let us use the result.

Suppose A is square and invertible, but we do not know its inverse. We want to find a square matrix B such that

$$BA = I$$

We can simply row-reduce A to the identity matrix using Gaussian elimination steps:

$$\underbrace{G_k G_{k-1} \cdots G_1}_{B} A = I$$

So

$$A^{-1} = B = G_k G_{k-1} \cdots G_1$$

A useful technique for keeping track of the multiplication of the elementary matrices G_i is to append the identity matrix to A, then do row reductions:

$$B \begin{bmatrix} A & I \end{bmatrix} = \begin{bmatrix} \underbrace{BA}_{I} & BI \end{bmatrix}$$

$$\begin{bmatrix} A & I \end{bmatrix} \to \begin{bmatrix} \underbrace{BA}_{I} & B \end{bmatrix}$$

Examples:

1. In field $\mathbb{F} = \mathbb{R}$, find A^{-1} for

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

Form an augmented matrix $[A \mid I]$ and do row reductions (the first step uses row 1 to modify row 3):

$$\left[\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 1 & 0 & 3 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 2 & 2 & 0 & 1 & 0 \\ 0 & -3 & -1 & -1 & 0 & 1 \end{array}\right] \quad \text{[now multiply row 2 by } 2^{-1}\text{]}$$

$$\to \left[\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1/2 & 0 \\ 0 & -3 & -1 & -1 & 0 & 1 \end{array}\right] \quad \text{[now use row 2 to modify row 3]}$$

$$\to \left[\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1/2 & 0 \\ 0 & 0 & 2 & -1 & 3/2 & 1 \end{array}\right] \quad \text{[now multiply row 3 by } 2^{-1}\text{]}$$

$$\to \left[\begin{array}{ccc|ccc} 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 1 & 1 & 0 & 1/2 & 0 \\ 0 & 0 & 1 & -1/2 & 3/4 & 1/2 \end{array}\right] \quad \text{[now use row 3 to modify rows 1,2]}$$

$$\to \left[\begin{array}{ccc|ccc} 1 & 3 & 0 & 3 & -3 & -2 \\ 0 & 1 & 0 & 1/2 & -1/4 & -1/2 \\ 0 & 0 & 1 & -1/2 & 3/4 & 1/2 \end{array}\right] \quad \text{[now use row 2 to modify row 1]}$$

$$\to \left[\begin{array}{ccc|ccc} 1 & 0 & 0 & 3/2 & -9/4 & -1/2 \\ 0 & 1 & 0 & 1/2 & -1/4 & -1/2 \\ 0 & 0 & 1 & -1/2 & 3/4 & 1/2 \end{array}\right]$$

So

$$A^{-1} = \begin{bmatrix} 3/2 & -9/4 & -1/2 \\ 1/2 & -1/4 & -1/2 \\ -1/2 & 3/4 & 1/2 \end{bmatrix}$$

We can verify the answer:

$$\begin{bmatrix} 3/2 & -9/4 & -1/2 \\ 1/2 & -1/4 & -1/2 \\ -1/2 & 3/4 & 1/2 \end{bmatrix}\begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 2 \\ 1 & 0 & 3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 1 & 3 & 4 \\ 0 & 2 & 2 \\ 1 & 0 & 3 \end{bmatrix}\begin{bmatrix} 3/2 & -9/4 & -1/2 \\ 1/2 & -1/4 & -1/2 \\ -1/2 & 3/4 & 1/2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

2. In field $\mathbb{F}_2 = \{0, 1\}$, find A^{-1} for

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

Do row reductions on $[A \mid I]$ (the first step uses row 1 to modify row 3):

$$\left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right] \to \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right]$$ [now swap rows 3 and 4]

$$\to \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{array}\right]$$ [now use row 4 to modify row 2]

$$\to \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{array}\right]$$ [now use row 4 to modify row 1]

$$\rightarrow \left[\begin{array}{cccc|cccc} 1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 \end{array}\right]$$

So

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

We can verify correctness:

$$\begin{bmatrix} 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{bmatrix} \underbrace{\begin{bmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}}_{I}$$

4.8.1. What if A is not invertible?

What happens if we use Gauss-Jordan on a square matrix that is not invertible? That is fine. Gauss-Jordan is a good way to determine whether or not a matrix is invertible. Matrix A is invertible if and only if it can be row-reduced to the identity (so there is a pivot on each row and column, with no free variables). If free variables appear then $Null(A) \neq \{0\}$ and hence A is not invertible. [The row operations needed to carry out Gauss-Jordan are the same as those for solving $Ax = 0$.]

The reader can use Gauss-Jordan to try finding A^{-1} on the following noninvertible matrix A with entries in $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ with mod 5 arithmetic:

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

In this case, the reader can "early-terminate" the procedure when row operations produce a row where the first three entries are all zero. The reader can then declare that A is not invertible. To officially do this, just go to a local coffee shop and proclaim to at least one barista: "A IS NOT INVERTIBLE."

4.8.2. A note on approximation error

When working with the fields $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, there is typically approximation error when solving $Ax = b$ or when inverting a square and invertible matrix A. For example, consider the following matrices A_1, A_2 and their inverses (recall Lemma 2.9):

$$A_1 = \begin{bmatrix} 1 & 1.001 \\ 1 & 1 \end{bmatrix} \implies A_1^{-1} = -1000 \begin{bmatrix} 1 & -1.001 \\ -1 & 1 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} 1 & 0.999 \\ 1 & 1 \end{bmatrix} \implies A_2^{-1} = 1000 \begin{bmatrix} 1 & -0.999 \\ -1 & 1 \end{bmatrix}$$

The slight perturbation between matrices A_1, A_2 produces a huge difference between A_1^{-1} and A_2^{-1} (the first entry of A_1^{-1} is -1000 while the first entry of A_2^{-1} is 1000). This type of *numerical instability* is something to watch out for. If A_1 was formed from measured data, we should be skeptical about whether or not it is really invertible: Perhaps the 1.001 entry is only due to measurement noise!

Numerical instability does not occur when working with finite fields. For example, suppose we work with $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ for p prime, say, $p = 2$ or $p = 5$ or $p = 103$. Then all entries of an invertible matrix A are in the set $\{0, 1, \ldots, p-1\}$, as are all entries of A^{-1}. Finite fields are ideal for finding clear and exact inverse matrices (when they exist) and for getting clear and exact solutions to $Ax = b$ (when solutions exist). The concept of *approximately equal to* does not exist for finite fields.

4.9. Bases for the fundamental subspaces of A

Let A be a nonzero $m \times n$ matrix. The Gaussian elimination steps to solve $Ax = 0$ give us r pivots and $n - r$ free variables (for some integer r). Since the number of pivots is less than or equal to the number of rows, and is also less than or equal to the number of columns, the value r must satisfy

$$r \in \{0, 1, \ldots, \min\{m, n\}\}$$

If there are no free variables then $Null(A) = \{0\}$. If there are free variables then we get a basis for $Null(A)$. This section shows that *with no extra work* we also get a basis for $Col(A)$ and for $Row(A)$ (where $Row(A) = Col(A^\top)$). That is, from just one Gaussian elimination procedure that row-reduces A to echelon form, we get bases for three out of the four fundamental subspaces of matrix A! (The remaining subspace $Null(A^\top)$ can be found by performing Gaussian elimination on $A^\top$.)

We first make precise claims, then we demonstrate with an example, then we explain why it works. Suppose we row reduce A to an echelon form matrix R, where each step

corresponds to left-multiplication by some elementary matrix G_i:

$$A \to \cdots \to R$$

$$G_k \cdots G_1 A = R$$

Suppose there are r pivot rows (these are the first r rows; they are nonzero and the rows underneath them are all zero). We already know how to find $Null(A)$ from R, and we know the dimension of $Null(A)$ is $n - r$ (the number of free variables).

Claim:

- The first r rows of R form a basis for $Row(A)$. [Formally, since $Row(A) = Col(A^\top)$, we should take the transpose of these r row vectors to turn them into r column vectors.]
- There are r pivot columns of R. While the columns of R in these pivot columns are *not necessarily* in $Col(A)$, the corresponding r columns of the *original matrix* A form a basis for $Col(A)$. That is, the pivot columns of R reveal the secret locations of columns in A that form a basis for $Col(A)$.

4.9.1. Example

Consider $\mathbb{F} = \mathbb{R}$ and

$$A = \begin{bmatrix} 1 & 2 & 3 & 0 & 7 \\ 0 & 1 & 1 & 3 & 2 \\ 2 & 0 & 2 & 0 & 5 \\ 2 & 1 & 3 & 4 & 6 \\ 0 & 0 & 0 & 0 & -1 \\ 1 & 3 & 4 & 3 & 9 \end{bmatrix}$$

Let us use Gaussian elimination to find solutions to $Ax = 0$. We shall stop early to declare bases for $Row(A)$ and $Col(A)$. We formally augment the matrix A with 0 because we are solving $Ax = 0$. We shall soon see this augmentation is not needed because that final zero column will never change. The first step uses row 1 to modify rows 3, 4, and 6:

$$\left[\begin{array}{ccccc|c} 1 & 2 & 3 & 0 & 7 & 0 \\ 0 & 1 & 1 & 3 & 2 & 0 \\ 2 & 0 & 2 & 0 & 5 & 0 \\ 2 & 1 & 3 & 4 & 6 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 1 & 3 & 4 & 3 & 9 & 0 \end{array}\right] \to \left[\begin{array}{ccccc|c} 1 & 2 & 3 & 0 & 7 & 0 \\ 0 & 1 & 1 & 3 & 2 & 0 \\ 0 & -4 & -4 & 0 & -9 & 0 \\ 0 & -3 & -3 & 4 & -8 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 1 & 1 & 3 & 2 & 0 \end{array}\right] \quad \text{[now use row 2 to modify rows 3,4,6]}$$

$$\rightarrow \left[\begin{array}{ccccc|c} 1 & 2 & 3 & 0 & 7 & 0 \\ 0 & 1 & 1 & 3 & 2 & 0 \\ 0 & 0 & 0 & 12 & -1 & 0 \\ 0 & 0 & 0 & 13 & -2 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \quad \text{[now use row 3 to modify row 4]}$$

$$\rightarrow \left[\begin{array}{ccccc|c} 1 & 2 & 3 & 0 & 7 & 0 \\ 0 & 1 & 1 & 3 & 2 & 0 \\ 0 & 0 & 0 & 12 & -1 & 0 \\ 0 & 0 & 0 & 0 & -11/12 & 0 \\ 0 & 0 & 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \quad \text{[now use row 4 to modify row 5]}$$

$$\rightarrow \underbrace{\left[\begin{array}{ccccc|c} \boxed{1} & 2 & 3 & 0 & 7 & 0 \\ 0 & \boxed{1} & 1 & 3 & 2 & 0 \\ 0 & 0 & 0 & \boxed{12} & -1 & 0 \\ 0 & 0 & 0 & 0 & \boxed{-11/12} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right]}_{R}$$

We now have echelon form. We could do back-substitution to get $Null(A)$. However, it is even easier to get $Row(A)$ and $Col(A)$. There are 4 pivots. The first four rows of our reduced matrix R give a basis for $Row(A)$ (the last 2 rows are all zero and clearly add nothing to the row space of A):

$$Row(A) = Col(A^\top) = Span\left\{\begin{bmatrix} 1 \\ 2 \\ 3 \\ 0 \\ 7 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 3 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 12 \\ -1 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ -11/12 \end{bmatrix}\right\} \tag{4.26}$$

The pivot columns have locations 1, 2, 4, 5. The corresponding columns of the original A matrix are a basis for $Col(A)$:

$$A = \begin{bmatrix} 1 & 2 & 3 & 0 & 7 \\ 0 & 1 & 1 & 3 & 2 \\ 2 & 0 & 2 & 0 & 5 \\ 2 & 1 & 3 & 4 & 6 \\ 0 & 0 & 0 & 0 & -1 \\ 1 & 3 & 4 & 3 & 9 \end{bmatrix} \Longrightarrow Col(A) = Span\left\{ \begin{bmatrix} 1 \\ 0 \\ 2 \\ 2 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 0 \\ 1 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 0 \\ 4 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 7 \\ 2 \\ 5 \\ 6 \\ -1 \\ 9 \end{bmatrix} \right\}$$

4.9.2. Why it works

Explanation of $Row(A)$

Recall that multiplying A on the left by an invertible matrix does not change the row space (Lemma 3.5). Since we are using row reductions, the relationship between A and R is

$$G_k \cdots G_1 A = R$$

So

$$Row(A) = Row(R) \tag{4.27}$$

By definition, $Row(R)$ is the span of the rows of R. The only nonzero rows of R are the first r rows (the remaining rows are all zero and add nothing to the span). So it is clear that these first r rows span $Row(R)$. However, the echelon form of these rows, where the first nonzero entry appears later and later as we go down the rows, ensures these rows are always linearly independent (see Exercise 4.14). Thus, in general, the first r rows of the echelon form matrix R form a basis for $Row(R)$, so (4.27) implies they are also a basis for $Row(A)$.

Explanation of $Col(A)$

LEMMA 4.2. *Let C, D be any matrices such that the multiplication CD makes sense. If the columns of CD are linearly independent, then the columns of D are linearly independent.*

PROOF. The proof of this deep result involves just one or two lines. You are invited to prove it yourself (see Exercise 3.16). [Hint: Recall that a matrix M has linearly independent columns if and only if $Null(M) = \{0\}$.] □

Our row reductions on A produced

$$GA = R$$

where G is some matrix (the fact that G is invertible plays no role in this particular argument). Matrix A has the same number of columns as matrix R. Writing this equation to emphasize the columns of A and R gives

$$G\underbrace{\begin{bmatrix} | & | & \cdots & | \\ a_1 & a_2 & \cdots & a_n \\ | & | & \cdots & | \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ r_1 & r_2 & \cdots & r_n \\ | & | & \cdots & | \end{bmatrix}}_{R}$$

where $a_1, \ldots, a_n$ are the columns of A; $r_1, \ldots, r_n$ are the columns of R. It follows that $Ga_i = r_i$ for $i \in \{1, \ldots, n\}$.

Now let $\tilde{R}$ be the matrix formed from R by taking only the r columns associated with pivots in matrix R (and throwing away the non-pivot columns). Let $\tilde{A}$ be the matrix formed from A by keeping the corresponding r columns of A. So we have

$$G\tilde{A} = \tilde{R}$$

It turns out that the pivot columns of R are always linearly independent: This is due to the pivot structure where the first pivot column has its first entry nonzero and all other entries zero; the second pivot column has its second entry nonzero and all entries below the second entry zero; and so on (see Exercise 4.14). So all columns of $\tilde{R}$ are linearly independent, but $\tilde{R} = G\tilde{A}$, so all columns of $G\tilde{A}$ are linearly independent, which means all columns of $\tilde{A}$ are linearly independent (by Lemma 4.2).

It remains to show the columns of $\tilde{A}$ also span $Col(A)$. These columns in $\tilde{A}$ are all columns of the original matrix A, so they are in $Col(A)$. There are exactly r of them, where r is the number of pivots and $n - r$ is the number of free variables (so $n - r$ is the dimension of $Null(A)$). A *rank-nullity theorem* proven in Chapter 5 implies that the dimension of $Col(A)$ plus the dimension of $Null(A)$ is always n, so r is the dimension of $Col(A)$. From the dimension theory to be developed in Chapter 5, we will learn that finding any r linearly independent vectors in a subspace of dimension r means that these vectors are a basis for the subspace.

4.9.3. A note on different ways to row reduce

There are many valid ways to take a nonzero matrix A and perform elementary row operations to reach an echelon form matrix R:

$$A \to \cdots \to R \quad , \quad G_k \cdots G_1 A = R$$

Different choices at each step may result in a different number of total steps, and may result in pivots in different locations. This also means that the free variables appear in different locations. This is not a problem: There is nothing fundamental about the locations of the pivots or the free variables. What *is* fundamental is the *number* of pivots: There are always r pivots (and hence $n - r$ free variables), where r is the dimension of $Col(A)$ and $n - r$ is the dimension of $Null(A)$.

If two equally valid row reduction methods lead to different pivot columns, it means that choosing the original columns of A either according to the pivot columns of method 1, or the pivot columns of method 2, are equally good for obtaining a basis for $Col(A)$. There can be many different ways to form a basis for $Col(A)$. Of course, every basis for $Col(A)$ has r vectors that are linearly independent and span $Col(A)$.

4.10. Solving $AY = B$

Suppose A and B are known matrices and we want to find a matrix Y such that

$$AY = B \tag{4.28}$$

Assume the sizes make sense for this multiplication:

$$A \sim m \times n$$
$$Y \sim n \times k$$
$$B \sim m \times k$$

Of course, if A is square and invertible then the one and only matrix Y that solves (4.28) is

$$Y = A^{-1}B$$

[For a formal (and easy) proof of this, see Exercise 4.15.]

What if A is not square, or is square but not necessarily invertible? Note that Y and B have the same number of columns, so let us write equation (4.28) in terms of those

columns:

$$A\underbrace{\begin{bmatrix} | & | & \cdots & | \\ y_1 & y_2 & \cdots & y_k \\ | & | & \cdots & | \end{bmatrix}}_{Y} = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ b_1 & b_2 & \cdots & b_k \\ | & | & \cdots & | \end{bmatrix}}_{B}$$

where $y_1, \ldots, y_k \in \mathbb{F}^n$ are the k columns of matrix Y; $b_1, \ldots, b_k \in \mathbb{F}^m$ are the k columns of B. This is equivalent to k separate equations:

$$Ay_i = b_i \quad \forall i \in \{1, \ldots, k\}$$

each of these is in the standard form $Ay = b$, so each can be separately solved by Gaussian elimination. However, doing this reuses a lot of the same row reduction steps. An interesting technique that only requires row reductions to be done *once* is to splice the entire B matrix onto A, then solve using standard Gaussian elimination on the augmented matrix.

For example, consider $\mathbb{F} = \mathbb{R}$ and

$$\begin{bmatrix} 1 & 1 & 4 & 0 \\ 0 & 1 & 1 & 1 \\ 3 & 0 & 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 & w_1 \\ x_2 & w_2 \\ x_3 & w_3 \\ x_4 & w_4 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 1 & 4 \\ 1 & 0 \end{bmatrix}$$

Then (with first step that uses row 1 to modify row 3):

$$\left[\begin{array}{cccc|cc} 1 & 1 & 4 & 0 & 1 & 3 \\ 0 & 1 & 1 & 1 & 1 & 4 \\ 3 & 0 & 1 & 1 & 1 & 0 \end{array}\right] \to \left[\begin{array}{cccc|cc} 1 & 1 & 4 & 0 & 1 & 3 \\ 0 & 1 & 1 & 1 & 1 & 4 \\ 0 & -3 & -11 & 1 & -2 & -9 \end{array}\right] \quad \text{[now use row 2 to modify row 3]}$$

$$\to \left[\begin{array}{cccc|cc} 1 & 1 & 4 & 0 & 1 & 3 \\ 0 & 1 & 1 & 1 & 1 & 4 \\ 0 & 0 & -8 & 4 & 1 & 3 \end{array}\right] \quad \text{[now multiply row 3 by } -1/8\text{]}$$

$$\to \left[\begin{array}{cccc|cc} 1 & 1 & 4 & 0 & 1 & 3 \\ 0 & 1 & 1 & 1 & 1 & 4 \\ 0 & 0 & 1 & -1/2 & -1/8 & -3/8 \end{array}\right] \quad \text{[now use row 3 to modify rows 1,2]}$$

$$\to \left[\begin{array}{cccc|cc} 1 & 1 & 0 & 2 & 3/2 & 9/2 \\ 0 & 1 & 0 & 3/2 & 9/8 & 35/8 \\ 0 & 0 & 1 & -1/2 & -1/8 & -3/8 \end{array}\right] \quad \text{[now use row 2 to modify row 1]}$$

$$\rightarrow \left[\begin{array}{cccc|cc} 1 & 0 & 0 & 1/2 & 3/8 & 1/8 \\ 0 & 1 & 0 & 3/2 & 9/8 & 35/8 \\ 0 & 0 & 1 & -1/2 & -1/8 & -3/8 \end{array}\right]$$

which means

$$x_1 + (1/2)x_4 = 3/8$$
$$x_2 + (3/2)x_4 = 9/8$$
$$x_3 + (-1/2)x_4 = -1/8$$

and

$$w_1 + (1/2)w_4 = 1/8$$
$$w_2 + (3/2)w_4 = 35/8$$
$$w_3 + (-1/2)w_4 = -3/8$$

So all possible solutions are

$$\begin{bmatrix} x_1 & w_1 \\ x_2 & w_2 \\ x_3 & w_3 \\ x_4 & w_4 \end{bmatrix} = \underbrace{\begin{bmatrix} 3/8 & 1/8 \\ 9/8 & 35/8 \\ -1/8 & -3/8 \\ 0 & 0 \end{bmatrix}}_{\text{particular solution } Y^*} + x_4 \begin{bmatrix} -1/2 & 0 \\ -3/2 & 0 \\ 1/2 & 0 \\ 1 & 0 \end{bmatrix} + w_4 \begin{bmatrix} 0 & -1/2 \\ 0 & -3/2 \\ 0 & 1/2 \\ 0 & 1 \end{bmatrix} \quad \forall x_4, w_4 \in \mathbb{R}$$

4.11. Convenient factorizations

Suppose we are given a matrix $A \sim m \times n$ with entries in some field $\mathbb{F}$. We want to solve *many* problems $Ax = b$ for different vectors $b \in \mathbb{F}^m$ (but with the same A). Rather than repeating the steps of Gaussian elimination again and again, we can put A into a nice form *once* and then use that form for speedy computations of $Ax = b$ every time we are given a new b.

If A is square and invertible, one way to do this is to compute the inverse matrix A^{-1}. While it takes time to compute A^{-1}, we only need to do it *once*. Then every time we are given a new $b \in \mathbb{F}^n$ we obtain the unique solution to $Ax = b$ by the multiplication

$$x = A^{-1}b$$

What if A is not square, or not invertible, or (even if it is invertible) we do not want to compute A^{-1}?

4.11.1. $GA = R$

Fix field $\mathbb{F}$. Let A have size $m \times n$ with entries in $\mathbb{F}$. Since we shall consider many different vectors b, for each $b \in \mathbb{F}^m$ define

$$X_b^* = \{x \in \mathbb{F}^n : Ax = b\}$$

We know that

$$X_b^* = x_b^* + Null(A)$$

where $Null(A)$ does not depend on b, while x_b^* is a particular solution to $Ax = b$ (which of course depends on b).

Recall that if G is any invertible $m \times m$ matrix then

$$X_b^* = \{x : Ax = b\} = \{x : GAx = Gb\} \quad \forall b \in \mathbb{F}^m \tag{4.29}$$

Using this, start with matrix A (suppose we do not yet have a b). Repeatedly multiply A on the left by elementary matrices G_i until the resulting product is in either basic echelon form or row reduced echelon form (as desired). Let R denote the reduced form of A:

$$\underbrace{G_k G_{k-1} \cdots G_1}_{G} A = R$$

Let G be the product of elementary matrices G_i that were used. Since G_i is invertible for each i, we know G is invertible and

$$\boxed{GA = R}$$

For an easy way to keep track of the running product of G_i matrices that form G we can do:

$$\left[\begin{array}{c|c} A & I \end{array}\right] \to \left[\begin{array}{c|c} R & G \end{array}\right] \qquad \Longleftrightarrow \qquad G_k G_{k-1} \cdots G_1 \left[\begin{array}{cc} A & I \end{array}\right] = \left[\begin{array}{cc} R & G \end{array}\right]$$

We now have a matrix G and a matrix R. Since G is invertible, we know

$$Null(A) = Null(R) \tag{4.30}$$

Since R is in echelon form, its null space is easy to find. So (4.30) gives us $Null(A)$ (either telling us $Null(A) = \{0\}$ if there are no free variables, or giving us a basis for $Null(A)$).

We do not yet have a vector b. However, from (4.29) we know for every $b \in \mathbb{F}^m$ that

$$X_b^* = \{x : Ax = b\} = \{x : GAx = Gb\} = \{x : Rx = Gb\}$$

which uses $GA = R$. Every time we are given a new $b \in \mathbb{F}^m$, we find a particular solution x_b^* (or declare that none exists) by using our precomputed R and G matrices to perform these two steps:

1. Left-multiply b by the matrix G. Call the resulting vector y. So $y = Gb$.
2. Find a particular solution x^* to $Rx = y$ by back-substitution (or declare that no solution exists). Since R is in echelon form, this is easily done by setting all free variables to zero.

The output is either a declaration that no solution exists (so $X_b^* = \phi$) or a particular solution x_b^* (so $X_b^* = x_b^* + Null(A)$).

4.11.2. $A = LU$ and $PA = LU$

Suppose we use Gaussian elimination to reduce a matrix A to basic echelon form R:

$$\underbrace{G_k \cdots G_1}_{G} A = R \tag{4.31}$$

Since we only require *basic* echelon form, all type 3 row operations can be done by adding a scalar multiple of a higher row into a lower row, corresponding to a *lower triangular* G_i matrix. Type 2 elementary matrices G_i are diagonal and hence lower triangular. Thus, if the row operations do not use any row swaps (a type 1 operation where G_i is a permutation matrix), then all elementary matrices G_i that are used are *lower triangular*. The inverse matrix G_i^{-1} for a type 2 or type 3 lower triangular matrix is again lower triangular (the explicit inverses are given in Sections 4.5.2-4.5.3). Also, the product of lower triangular matrices is again lower triangular (see Exercise 2.21 and Chapter 2.10.1):

$$G = G_k G_{k-1} \cdots G_1 = \text{lower triangular}$$
$$G^{-1} = G_1^{-1} \cdots G_{k-1}^{-1} G_k^{-1} = \text{lower triangular}$$

Define

$$L = G^{-1} = \text{lower triangular}$$

where L emphasizes *lower triangular*. Then from (4.31)

$$GA = R \implies A = G^{-1}R = LR$$

Now suppose the original matrix A was square. It means the echelon form matrix R is also square, and a square echelon form matrix must be *upper triangular*. So define $U = R$, where U stands for *upper triangular*:

$$A = LR = LU$$

Thus, if we can row reduce a square matrix A to an upper triangular matrix U *without performing any row swap operations*, then we have a famous factorization:

$$A = LU$$

which factors A into the product of an invertible lower triangular matrix L and a (possibly noninvertible) upper triangular matrix U. Such a factorization can also be used for speedy computations of $Ax = b$, reusing the same factorization $A = LU$ for each new b. Since $Ax = b$ is the same as $LUx = b$, we do

1. Find a vector y that solves $Ly = b$. Since L is a lower triangular matrix, this is easily accomplished by back-substitution. [A solution to $Ly = b$ always exists because L is invertible.]
2. Find a vector x that solves $Ux = y$ (using back-substitution, since U is upper triangular). [If U is not invertible, we may or may not need to declare that no solution exists.]

It should be emphasized that not all square matrices A can be factored into $A = LU$ (see Exercise 4.42 for a simple example). We are only guaranteed such a factorization exists if A can be reduced to an upper triangular matrix using some sequence of row operations that *does not involve row swaps*. Fortunately, we can modify the factorization result: Suppose that through elementary row operations we have something like

$$L_5P_3L_4P_2L_3P_1L_2L_1A = U \tag{4.32}$$

where L_i stand for lower triangular matrices, and P_i stand for permutation (row swap) matrices. It can be shown that if we *push all the permutation matrices to the front* and then perform row operations, we never need any more row swaps. So there are type 2 and 3 elementary matrices $\tilde{L}_1, \ldots, \tilde{L}_k$ that are lower triangular and

$$\underbrace{\tilde{L}_k\tilde{L}_{k-1}\cdots\tilde{L}_1}_{\text{lower triangular}} \cdot \underbrace{P_3P_2P_1}_{P} A = U \tag{4.33}$$

This is remarkable because the L_i and P_i matrices in (4.32) do not necessarily commute.

Define

$$\begin{aligned} P &= P_3 P_2 P_1 \\ \tilde{L} &= \tilde{L}_k \tilde{L}_{k-1} \cdots \tilde{L}_1 = \text{lower triangular} \\ L &= \tilde{L}^{-1} = \tilde{L}_1^{-1} \cdots \tilde{L}_{k-1}^{-1} \tilde{L}_k^{-1} = \text{lower triangular} \end{aligned}$$

Then (4.33) becomes

$$\tilde{L}PA = U$$

Multiplying both sides on the left by L and using $L\tilde{L} = I$ gives

$$\boxed{PA = LU}$$

Summary: Any square matrix A can be factored so that $PA = LU$ where P is some permutation matrix, L is some invertible lower triangular matrix, and U is some (not necessarily invertible) upper triangular matrix.

Corresponding facts: The above line of reasoning suggests a good proof of the following facts:

- Every invertible lower triangular matrix A has a lower triangular inverse.
- Every invertible upper triangular matrix A has an upper triangular inverse.

The second bullet can be proven from the first by taking transposes. To prove the first: It can be shown that all diagonal entries a_{ii} of an invertible lower triangular matrix A must be nonzero (else $a_{ii} = 0$ and a back-substitution step of the type (4.6) will fail to find a solution to $Ax = b$ for all vectors b). These diagonals are then already in place as pivots. So A can be row reduced to the identity without using row swaps via lower triangular G_i matrices:

$$(G_k G_{k-1} \cdots G_1 A = I) \implies \left(A^{-1} = G_k G_{k-1} \cdots G_1 = \text{lower triangular}\right)$$

where we use the fact that a product of lower triangulars is lower triangular.

4.12. Exercises

EXERCISE 4.1. *(What is wrong with this logic?) What is incorrect about the following logic?*

Using $\mathbb{F} = \mathbb{R}$, we want to find real numbers x and y to solve:

$$\begin{bmatrix} 1 & 1 \\ -2 & 3 \\ 6 & -10 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

Left-multiplying both sides of the equation by the same matrix gives:

$$\underbrace{\begin{bmatrix} 1 & 3 & 1 \\ 0 & -3 & -1 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ -2 & 3 \\ 6 & -10 \end{bmatrix}}_{I} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 1 & 3 & 1 \\ 0 & -3 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}$$

where I is the 2×2 identity matrix. Simplifying the left-hand-side and right-hand-side gives

$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 8 \\ -7 \end{bmatrix}$$

therefore $(x, y) = (8, -7)$.

EXERCISE 4.2. *(Solution set of an equation) Assume A, B are matrices of sizes appropriate so that BA makes sense. Let b be a column vector with the same number of rows as A. What can be learned from Exercise 4.1 about the solution sets $\{x : Ax = b\}$ and $\{x : BAx = Bb\}$?*

EXERCISE 4.3. *(The solution set X^* is affine) Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Fix $b \in \mathbb{F}^m$ and define $X^* = \{x \in \mathbb{F}^n : Ax = b\}$. Show that for any $k \in \mathbb{N}$, any vectors $v_1, ..., v_k \in X^*$, and any scalars $a_1, ..., a_k$ that satisfy $\sum_{i=1}^k a_i = 1$ we have*

$$\sum_{i=1}^{k} a_i v_i \in X^*$$

[The linear combination $\sum_{i=1}^k a_i v_i$ is called an affine combination *because the scalar coefficients sum to 1. The property shown in this exercise means that X^* is an* affine set. *Affine sets are treated for general vector spaces in Appendix C.]*

EXERCISE 4.4. *Consider $\mathbb{F}_7 = \{0, 1, 2, 3, 4, 5, 6\}$ with mod 7 arithmetic. Suppose A has size 100×87 and has entries in $\mathbb{F}_7$. Suppose $\{v_1, ..., v_{38}\}$ is a basis for Null(A).*

a) Evaluate $A(5v_3 + 2v_9)$.

b) Are the columns of A linearly independent?

c) Suppose $b \in Col(A)$. How many solutions are there to $Ax = b$?

d) Suppose $b \notin Col(A)$. How many solutions are there to $Ax = b$?

EXERCISE 4.5. *Consider $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic. Suppose A has size 1000×1000 and has entries in $\mathbb{F}_3$. Suppose $\{v_1, ..., v_{20}\}$ is a basis for $Null(A)$.*

a) Evaluate $A(v_{19} - 2v_{20})$.

b) Are the columns of A linearly independent?

c) Suppose $b \in Col(A)$. How many solutions are there to $Ax = b$?

d) Suppose $b \notin Col(A)$. How many solutions are there to $Ax = b$?

EXERCISE 4.6. *We use the binary field $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic. Our matrix A has size 27×25 and has entries in $\mathbb{F}_2$. The columns of A are linearly independent.*

a) Suppose $b \in Col(A)$. How many solutions are there to $Ax = b$?

b) Suppose $b \notin Col(A)$. How many solutions are there to $Ax = b$?

c) Is A invertible?

EXERCISE 4.7. *We use the finite field $\mathbb{F}_{17}$ with mod 17 arithmetic. Our matrix A has size 1038×512 and has entries in $\mathbb{F}_{17}$. The columns of A are linearly independent.*

a) Suppose $b \in Col(A)$. How many solutions are there to $Ax = b$?

b) Suppose $b \notin Col(A)$. How many solutions are there to $Ax = b$?

c) Is A invertible?

EXERCISE 4.8. *Give an example of a field and a matrix A and vector $b \neq 0$ with entries in that field such that $Ax = b$ has exactly 125 solutions.*

EXERCISE 4.9. *Give an example of a field and a matrix A and vector b with entries in that field such that $Ax = b$ has no solution, $Null(A)$ has 49 vectors, and $Col(A)$ has more than one vector.*

EXERCISE 4.10. *Give an example of a field and a matrix A and vector $b \neq 0$ with entries in that field such that $Ax = b$ has exactly 32 solutions.*

EXERCISE 4.11. *Give an example of a field and a matrix A and vector b with entries in that field such that $Ax = b$ has no solution, $Null(A)$ has 256 vectors, and $Col(A)$ has 4 vectors.*

EXERCISE 4.12. *Consider the vectors $x^*, z_1, z_2, z_3, z_4 \in \mathbb{R}^6$ given in the first example of Section 4.4.1. The vector x^* is defined in terms of a vector $c = (c_1, c_2, 0, 0) \in \mathbb{R}^4$. Let R be the matrix given there. We know that*

$$R\left(x^* + \sum_{i=1}^{4} a_i z_i\right) = c \quad \forall a_1, a_2, a_3, a_4 \in \mathbb{R} \tag{4.34}$$

and it was shown in that section that all solutions to $Rx = c$ are given by $x^* + Span\{z_1, z_2, z_3, z_4\}$.

a) Use (4.34) *to prove that* $Rx^* = c$.

b) Use (4.34) *to prove* $z_i \in Null(R)$ *for all* $i \in \{1, 2, 3, 4\}$.

c) Using the particular numerical values of $\{z_1, z_2, z_3, z_4\}$ *given in that section, prove these four vectors are linearly independent. [For example, the particular numerical values for* z_1 *and* z_2 *are* $z_1 = (1, 0, 0, 0, 0, 0)$ *and* $z_2 = (0, -1/2, 1, 0, 0, 0)$*.]*

d) How can it be argued that $Null(R) = Span\{z_1, z_2, z_3, z_4\}$*? [Hint: Repeat the argument of the first example in Section 4.4.1 for the special case* $c_1 = c_2 = 0$*.]*

EXERCISE 4.13. *(A binary matrix in* $\mathbb{R}$ *and* $\mathbb{F}_2$*) Consider*

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

a) Consider field $\mathbb{F} = \mathbb{R}$*. Determine if A is invertible. If so, compute* A^{-1} *using the Gauss-Jordan method.*

b) Now consider A as a matrix with entries in the binary field $\mathbb{F}_2 = \{0, 1\}$ *with mod 2 arithmetic. Determine if A is invertible. If so, compute* A^{-1} *using the Gauss-Jordan method.*

EXERCISE 4.14. *(Pivot rows of R are linearly independent) Let* $\{r_1, r_2, r_3, r_4\}$ *be the four pivot rows from* (4.26) *(they are row vectors when in R, but we treat them as column vectors now). So* $r_1 = (1, 2, 3, 0, 7)$*,* $r_2 = (0, 1, 1, 3, 2)$*, and so on*

a) Prove $\{r_1, r_2, r_3, r_4\}$ *are linearly independent.*

b) In general, suppose we have a collection of vectors $\{v_1, ..., v_k\}$ *in* $\mathbb{F}^n$ *such that for all* $i \in \{1, ..., k\}$*, vector* v_i *has a nonzero entry in some particular row m, while all vectors* v_j *for* $j > i$ *have a zero entry in row m. Show* $\{v_1, ..., v_k\}$ *are linearly independent.*

EXERCISE 4.15. *Suppose* $A \sim m \times n$ *and* $B \sim m \times k$ *are known matrices. We want to find a matrix Y that solves*

$$AY = B \tag{4.35}$$

a) What is the required size of matrix Y?

b) Suppose A is invertible and define $Y^* = A^{-1}B$*. Show that* Y^* *is a solution to* (4.35)*.*

c) Continue to assume A is invertible. Show that the matrix Y^* *of part (b) is the one and only solution to* (4.35)*. That is, when A is invertible, the solution to* (4.35) *exists and is unique.*

d) Give an example of matrices A, B for which there is more than one solution Y to (4.35)*.*

e) Give an example of matrices A, B for which there is no solution matrix Y to (4.35)*.*

EXERCISE 4.16. *Using the mod-2 field* $\mathbb{F}_2 = \{0, 1\}$*, find a* 3×3 *matrix B such that* $AB = C$*, where A and C are given below:*

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

EXERCISE 4.17. *(Elementary row operation) We have a* 3×3 *matrix* A *with entries in some field* $\mathbb{F}$*:*

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

a) Fix $\theta \in \mathbb{F}$*. Give a* 3×3 *matrix* G *such that*

$$GA = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ (\theta a_{11} + a_{31}) & (\theta a_{12} + a_{32}) & (\theta a_{13} + a_{33}) \end{bmatrix}$$

b) What is G^{-1}*?*

EXERCISE 4.18. *(Matrix equation) Solve the matrix equation* $A\vec{x} = \vec{b}$ *shown below, where all arithmetic is performed in* $\mathbb{F}_3 = \{0, 1, 2\}$*. Thus, all arithmetic is taken mod 3, and the values* x*,* y*, and* z *are in* $\mathbb{F}_3 = \{0, 1, 2\}$*.*

$$\begin{bmatrix} 0 & 1 & 2 \\ 1 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ 2 \end{bmatrix}$$

EXERCISE 4.19. *(Basis for Null*(A)*) Find a basis for the nullspace of* A *when* A *is the following real matrix:*

$$A = \begin{bmatrix} 0 & 1 & -1 & 2 \\ 2 & -1 & 5 & -6 \\ 2 & 3 & 1 & 2 \\ 3 & 0 & 6 & -6 \end{bmatrix}$$

EXERCISE 4.20. *Fix* $\mathbb{F} = \mathbb{R}$*. Give an example of matrices* A, C *and a vector* b *(of appropriate size) such that the equation* $Ax = b$ *has strictly fewer solutions than the equation* $CAx = Cb$*.*

EXERCISE 4.21. *(Are they linearly independent?) Suppose* $\{v_1, v_2, v_3\}$ *are linearly independent vectors. Assume the field* $\mathbb{F} = \mathbb{R}$*.*

a) Are $\{5v_1 - v_2, v_1 + v_2\}$ *linearly independent vectors?*

b) Are $\{6v_1 - 2v_2, -3v_1 + v_2\}$ *linearly independent vectors?*

c) Are $\{8v_1 + v_3, -v_2 + 4v_3, v_1 + v_2\}$ *linearly independent vectors?*

d) Are $\{v_1 + 2v_2 - 2v_3, -v_2 + 3v_3, v_1 + 7v_2 - 17v_3\}$ *linearly independent?*

EXERCISE 4.22. *Suppose* $\{x_1, ..., x_k\}$ *are linearly independent vectors.*

a) Fix $j \in \{1, ..., k\}$*. Show that* $\{x_1, ..., x_j\}$ *are linearly independent.*

b) Let $g : \{1, ..., k\} \to \{1, ..., k\}$ *be a bijective function (a one-to-one correspondence). Thus,* $(g[1], g[2], ..., g[k])$ *is a permutation of the indices* $(1, 2, ..., k)$*. Show that* $\{x_{g[1]}, x_{g[2]}, ..., x_{g[k]}\}$ *are linearly independent.*

EXERCISE 4.23. *(Set of solutions) Consider the binary field* $\mathbb{F}_2 = \{0, 1\}$ *with mod 2 arithmetic. Consider*

$$\begin{bmatrix} 0 & 1 & 1 & 0 & 1 \\ 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}$$

a) What is the set X^* *of all solutions? What is* $|X^*|$ *(the number of elements in* X^**)?*

b) Find $Null(A)$*. What is the relationship between* X^* *and* $Null(A)$*?*

EXERCISE 4.24. *Let* $\mathbb{F}$ *be some field. Let* A *be a* $m \times n$ *matrix. Let* G *be an invertible* $n \times n$ *matrix. Prove that* $Col(A) = Col(AG)$*. Use the technique of first showing* $Col(A) \subseteq Col(AG)$ *and then showing* $Col(AG) \subseteq Col(A)$*.*

EXERCISE 4.25. *(Finding an unknown matrix) Suppose* $f : \mathbb{R}^2 \to \mathbb{R}^2$ *is the following nonlinear function*

$$f(x_1, x_2) = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} x_1^2 \\ x_1 + x_2 \end{bmatrix} \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

where A *is an unknown* 2×2 *matrix with real entries. We test two input vectors to obtain the following information:*

$$f(1, 1) = \begin{bmatrix} 5 \\ 3 \end{bmatrix} \quad , \quad f(2, -1) = \begin{bmatrix} 8 \\ 0 \end{bmatrix}$$

What is the matrix A*?*

EXERCISE 4.26. *Consider the* 3×3 *matrix* A *given by*

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

a) Using the field $\mathbb{R}$*, find a basis for* $Null(A)$*,* $Row(A)$*, and* $Col(A)$*.*

b) Using the binary field $\mathbb{F}_2 = \{0, 1\}$*, find a basis for* $Null(A)$*,* $Row(A)$*, and* $Col(A)$*.*

EXERCISE 4.27. *Consider the* 4×4 *matrix* A *given by*

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

Using the binary field $\mathbb{F}_2 = \{0, 1\}$*, find a basis for* $Null(A)$*,* $Col(A)$*,* $Null(A^\top)$*,* $Col(A^\top)$*.*

EXERCISE 4.28. *Consider the* 3×4 *matrix* A *given by*

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

a) Using the field $\mathbb{R}$*, find a basis for* $Null(A)$ *and* $Col(A)$*.*

b) Using the binary field $\mathbb{F}_2 = \{0, 1\}$*, find a basis for* $Null(A)$ *and* $Col(A)$*.*

EXERCISE 4.29. *Using field* $\mathbb{R}$*, find a basis for* $Null(A)$*,* $Col(A)$*,* $Null(A^\top)$*,* $Col(A^\top)$ *when*

$$A = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 1 & 7 \\ -1 & 3 & 11 \end{bmatrix}$$

EXERCISE 4.30. *Using field* $\mathbb{R}$*, find a basis for* $Null(A)$*,* $Col(A)$*,* $Null(A^\top)$*,* $Col(A^\top)$ *when*

$$A = \begin{bmatrix} 0 & 1 & 2 & 1 \\ -1 & 1 & 7 & 0 \\ -1 & 3 & 11 & 2 \end{bmatrix}$$

EXERCISE 4.31. *Using field $\mathbb{R}$, find a basis for $Null(A)$, $Col(A)$, $Null(A^\top)$, $Col(A^\top)$ when*

$$A = \begin{bmatrix} 8 & 4 & 0 & -1 \\ 0 & 5 & 7 & 0 \\ 0 & 0 & 1 & 2 \end{bmatrix}$$

EXERCISE 4.32. *Using field $\mathbb{F}_2 = \{0, 1\}$, find a basis for $Null(A)$, $Col(A)$, $Null(A^\top)$, $Col(A^\top)$ when*

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

EXERCISE 4.33. *Using field $\mathbb{F}_2 = \{0, 1\}$, find a basis for $Null(A)$, $Col(A)$, $Null(A^\top)$, $Col(A^\top)$ when*

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

EXERCISE 4.34. *Using $\mathbb{F} = \mathbb{R}$, find two different bases for $Row(A)$ when*

$$A = \begin{bmatrix} 1 & 2 & 3 & 0 \\ 2 & 4 & 6 & 0 \\ 0 & 2 & -1 & 1 \\ 4 & 2 & 15 & -3 \\ -7 & -2 & -27 & 6 \end{bmatrix}$$

EXERCISE 4.35. *(Inverting matrices) The following can be solved by the Gauss-Jordan method.*

a) Using $\mathbb{F} = \mathbb{R}$ with standard arithmetic, find A^{-1}:

$$A = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 2 & 0 \\ 0 & -2 & 8 \end{bmatrix}$$

b) Using $\mathbb{F} = \{0, 1\}$ with mod 2 arithmetic, find A^{-1}:

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

EXERCISE 4.36. *(Inverting a matrix with entries in $\mathbb{F}_3$) Using $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic, find A^{-1} when*

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 1 & 1 & 0 \\ 0 & 2 & 1 \end{bmatrix}$$

EXERCISE 4.37. *(Working with complex numbers) Using $\mathbb{F} = \mathbb{C}$, find a basis for $Null(A)$ and $Row(A)$ when A is the following 3×3 matrix:*

$$A = \begin{bmatrix} i & 1 & 2 \\ 13 + i & 4 + 2i & -1 - 2i \\ 4 & 1 + i & -1 \end{bmatrix}$$

EXERCISE 4.38. *Define the following vectors x, y, z in $\mathbb{R}^3$:*

$$x = \begin{bmatrix} 1 \\ 2 \\ 2 \end{bmatrix} \quad , \quad y = \begin{bmatrix} 5 \\ -1 \\ 0 \end{bmatrix} \quad , \quad z = \begin{bmatrix} 3 \\ 3 \\ 3 \end{bmatrix}$$

a) Show that $\{x, y, z\}$ are linearly independent.

b) Express $(9, 8, 7) \in \mathbb{R}^3$ as a linear combination of $\{x, y, z\}$.

c) Express $(-7, 0, 7) \in \mathbb{R}^3$ as a linear combination of $\{x, y, z\}$.

EXERCISE 4.39. *Define the following vectors x, y, z in $\mathbb{R}^3$:*

$$x = \begin{bmatrix} 0 \\ 9 \\ 9 \end{bmatrix} \quad , \quad y = \begin{bmatrix} 2 \\ 0 \\ 2 \end{bmatrix} \quad , \quad z = \begin{bmatrix} 3 \\ 3 \\ 1 \end{bmatrix}$$

a) Show that $\{x, y, z\}$ are linearly independent.

b) Express $(9, 8, 7) \in \mathbb{R}^3$ as a linear combination of $\{x, y, z\}$.

c) Express $(-7, 0, 7) \in \mathbb{R}^3$ as a linear combination of $\{x, y, z\}$.

EXERCISE 4.40. *Define the following vectors x, y, z, w in $\mathbb{R}^4$*

$$x = \begin{bmatrix} 2 \\ 2 \\ 0 \\ 0 \end{bmatrix}, \quad y = \begin{bmatrix} 3 \\ 0 \\ 3 \\ 0 \end{bmatrix}, \quad z = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad w = \begin{bmatrix} -1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

a) Show that $\{x, y, z, w\}$ are linearly independent.
b) Express $(12, 12, 18, 6) \in \mathbb{R}^4$ as a linear combination of $\{x, y, z, w\}$.
c) Express $(7, 0, 0, 7) \in \mathbb{R}^4$ as a linear combination of $\{x, y, z, w\}$.

EXERCISE 4.41. *Define the following vectors x, y, z, w in $\mathbb{R}^4$*

$$x = \begin{bmatrix} 1 \\ 2 \\ 0 \\ 0 \end{bmatrix}, \quad y = \begin{bmatrix} -3 \\ 1 \\ 3 \\ 0 \end{bmatrix}, \quad z = \begin{bmatrix} 2 \\ 1 \\ 2 \\ 1 \end{bmatrix}, \quad w = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}$$

a) Show that $\{x, y, z, w\}$ are linearly independent.
b) Express $(12, 12, 18, 6) \in \mathbb{R}^4$ as a linear combination of $\{x, y, z, w\}$.
c) Express $(7, 0, 0, 7) \in \mathbb{R}^4$ as a linear combination of $\{x, y, z, w\}$.

EXERCISE 4.42. *(A square matrix with no LU factorization) For $\mathbb{F} = \mathbb{R}$ define*

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

Prove there is no way to write $A = LU$ where L is a lower triangular 2×2 matrix and U is an upper triangular 2×2 matrix, both with entries in $\mathbb{R}$. For this, suppose there are matrices of the form

$$L = \begin{bmatrix} a & 0 \\ b & c \end{bmatrix}, \quad U = \begin{bmatrix} d & e \\ 0 & f \end{bmatrix}$$

for some real numbers a, b, c, d, e, f, for which $A = LU$. Reach a contradiction. (In fact, this same matrix A works as a counter-example in any field $\mathbb{F}$.)

CHAPTER 5.

Dimension and rank

This chapter establishes the concept of *dimension* for a subspace of $\mathbb{F}^n$. It defines the *rank* of a matrix A as the dimension of $Col(A)$. It proves that the rank of A, plus the dimension of $Null(A)$, is equal to the number of columns of A (the rank-nullity theorem). These definitions and results do not require the Gaussian elimination algorithm. However, they are consistent with the treatment in Chapter 4 that defines the rank of a matrix A to be the number of pivots produced by Gaussian elimination. That is because row-reducing A to echelon form always creates a number of pivots equal to the dimension of $Col(A)$ and a number of free variables equal to the dimension of $Null(A)$.

5.1. Conditions for linear dependence and independence

Exercise 4.22 shows that rearranging the order of linearly independent vectors does not change linear independence. For example, if $\{x, y, z\}$ are linearly independent, then so are $\{y, z, x\}$. Similarly, rearranging the order of linearly *dependent* vectors maintains their linear dependence. Exercise 3.21 shows that vectors $\{x_1, ..., x_k\}$ are linearly *dependent* if any one of these two *sufficient conditions* holds:

- The zero vector is in the collection: There is an index $i \in \{1, ..., k\}$ such that $x_i = 0$.
- Two vectors are repeated: There are indices $i, j \in \{1, ..., k\}$ such that $i \neq j$ and $x_i = x_j$.

The following lemmas are more general and provide conditions that are both necessary and sufficient.

LEMMA 5.1. *Vectors $\{x_1, ..., x_k\}$ are linearly dependent if and only if either $x_1 = 0$, or there is a $j \in \{2, ..., k\}$ such that*

$$x_j \in Span\{x_1, ..., x_{j-1}\}$$

LEMMA 5.2. *Vectors $\{x_1, ..., x_k\}$ are linearly independent if and only if all of the following hold:*

- $x_1 \neq 0$.
- $x_2 \notin Span\{x_1\}$.
- $x_3 \notin Span\{x_1, x_2\}$.
- $x_4 \notin Span\{x_1, x_2, x_3\}$.
- ...
- $x_k \notin Span\{x_1, ..., x_{k-1}\}$.

Lemma 5.1 is interesting because we know that linear dependence of $\{x_1, ..., x_k\}$ does not depend on the *ordering* of these vectors. Thus, if these vectors are linearly dependent, then for any different ordering of these vectors, either the first vector in the new ordering is 0, or there is some vector that is a linear combination of its predecessors within the new ordering. Lemma 5.2 is an immediate consequence of Lemma 5.1 (it is the contrapositive statement). Below we provide a proof of Lemma 5.1.

PROOF. (Lemma 5.1) First suppose $\{x_1, ..., x_k\}$ are linearly dependent. Then there are scalars $a_1, \ldots, a_k$, not all zero, such that

$$\sum_{i=1}^{k} a_i x_i = 0$$

Let j be the largest index in $\{1, ..., k\}$ for which $a_j \neq 0$ (such an index must exist because not all of the scalars are zero). Then $a_i = 0$ for $i > j$ and the above equation reduces to

$$\sum_{i=1}^{j} a_i x_i = 0$$

where $a_j \neq 0$. If $j = 1$ then the equation becomes $a_1 x_1 = 0$, and since $a_1 \neq 0$ this means $x_1 = 0$

and we are done. If $j > 1$ then we have

$$\left(\sum_{i=1}^{j-1} a_i x_i\right) + a_j x_j = 0$$

and since $a_j \neq 0$ we can rearrange this to give

$$x_j = \sum_{i=1}^{j-1} (-a_i/a_j) x_i$$

and so x_j is a linear combination of its predecessors.

To prove the other direction, suppose that either $x_1 = 0$, or there is an index $j \in \{2, \ldots, k\}$ such that x_j is a linear combination of $\{x_1, \ldots, x_{j-1}\}$. If $x_1 = 0$ then

$$1 \cdot x_1 + \sum_{j=2}^{k} 0 x_j = 0$$

and this is a linear combination that gives zero, but not all the scalar weights are zero. So $\{x_1, \ldots, x_k\}$ are linearly dependent. If $x_1 \neq 0$ then there is a $j \geq 2$ such that $x_j = \sum_{i=1}^{j-1} a_i x_i$ for some scalars $a_1, \ldots, a_{j-1}$. Then

$$\left(\sum_{i=1}^{j-1} a_i x_i\right) + (-1) x_j + \sum_{i=j+1}^{k} (0) x_i = 0$$

and this is a linear combination that gives zero, even though not all scalar weights are zero (since -1 is one of the scalar weights). So again $\{x_1, \ldots, x_k\}$ are linearly dependent. □

5.2. Wide matrix theorem

Let $A \sim m \times n$ be a matrix with entries in some field $\mathbb{F}$.

THEOREM 5.1. *(Wide matrix theorem) Every wide matrix has a nontrivial null space.*

This is a foundational theorem of linear algebra because it can be used to establish a concept of *dimension* for subspaces of $\mathbb{F}^n$ and for more general vector spaces treated in Chapter 6. Theorem 5.1 can be proven using the procedure of Gaussian elimination to find all solutions to $Ax = 0$. Gaussian elimination produces an echelon form matrix with at most one pivot on each row. Therefore, if the number of columns is larger than the number of rows, there must be at least one column with no pivot. That column allows a

free variable, so $Null(A)$ is nontrivial. An alternative proof of the wide matrix theorem that does not use the Gaussian elimination algorithm is given in Exercises 5.45-5.46. That proof uses a *vector replacement* technique that is similar to the treatment in Axler (2024).

The first fundamental consequence of Theorem 5.1 is this: If we have a collection of n vectors in $\mathbb{F}^m$, we can put them together as columns to form a $m \times n$ matrix A. If $n > m$ then A is wide, so Theorem 5.1 implies $Null(A) \neq \{0\}$ and hence the columns of A are linearly dependent. Therefore, if $\{x_1, \ldots, x_n\}$ is a collection of *linearly independent* vectors in $\mathbb{F}^m$, then we must have $n \leq m$. That is, it is impossible to have a collection of more than m linearly independent vectors in the vector space $\mathbb{F}^m$. The next theorem, called the independent-set-spanning-set comparison theorem, is a generalization of this idea. We first recall a definition: We say that vectors $\{y_1, ..., y_m\}$ *span a subspace* S if $Span\{y_1, ..., y_m\} = S$.

THEOREM 5.2. *(Independent-set-spanning-set comparison) Let S be a subspace of $\mathbb{F}^n$. Suppose $x_1, \ldots, x_k, y_1, \ldots, y_m$ are vectors in S such that*

- *$\{x_1, \ldots, x_k\}$ are linearly independent in S.*
- *$\{y_1, \ldots, y_m\}$ span S.*

Then $k \leq m$.

PROOF. Suppose $k > m$ (we reach a contradiction). For each $j \in \{1, ..., k\}$, since $x_j \in S$ and the vectors $\{y_1, ..., y_m\}$ span S, the vector x_j must be a linear combination of $\{y_1, \ldots, y_m\}$:

$$x_j = \sum_{i=1}^{m} a_{ij} y_i \quad \text{(for some scalars } a_{ij})$$

This can be written in matrix form

$$x_j = \begin{bmatrix} | & | & \cdots & | \\ y_1 & y_2 & \cdots & y_m \\ | & | & \cdots & | \end{bmatrix} \begin{bmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{bmatrix} \quad \forall\, j \in \{1, ..., k\}$$

From this, we can stack all the column vectors $x_1, ..., x_k$ together in matrix form:

$$\underbrace{\begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_k \\ | & | & \cdots & | \end{bmatrix}}_{X} = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ y_1 & y_2 & \cdots & y_m \\ | & | & \cdots & | \end{bmatrix}}_{Y} \underbrace{\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1k} \\ a_{21} & a_{22} & \cdots & a_{2k} \\ \vdots & & & \\ a_{m1} & a_{m2} & \cdots & a_{mk} \end{bmatrix}}_{A}$$

As illustrated by the underbraces in the above equation: define X as the matrix with columns x_j; define Y as the matrix with columns y_i; define $A = (a_{ij})$. So

$$X = YA \tag{5.1}$$

Matrix A has size $m \times k$. Since $k > m$, matrix A is wide. Theorem 5.1 implies $Null(A) \neq \{0\}$. Thus, there is a vector $z \neq 0$ such that $Az = 0$. Multiplying (5.1) on the right by z gives

$$Xz = YAz = 0$$

Since $Xz = 0$ but $z \neq 0$, the columns of X cannot be linearly independent, contradicting the fact that $\{x_1, ..., x_k\}$ are linearly independent. □

5.3. Basis and dimension

The definition of *basis for a subspace* was given in Chapter 3.9. Recall that if S is a subspace of $\mathbb{F}^n$, then a collection of vectors $\{x_1, \ldots, x_k\}$ in S is said to be a *basis for* S if:

1. $\{x_1, ..., x_k\}$ are linearly independent.
2. $Span\{x_1, ..., x_k\} = S$

The full vector space $\mathbb{F}^n$ has the *standard basis* $\{e_1, e_2, ..., e_n\}$, where e_i is the ith column of the $n \times n$ identity matrix (recall Lemma 3.11). By comparing any linearly independent collection of vectors in $\mathbb{F}^n$ with the spanning set $\{e_1, ..., e_n\}$, the independent-set-spanning-set comparison theorem ensures the linearly independent collection has at most n vectors. [Of course, the wide matrix theorem also ensures this.]

The independent-set-spanning-set comparison theorem makes it easy to show any two bases for the same subspace must contain the same number of vectors. Indeed, let S be a subspace and let $\{x_1, ..., x_k\}$ and $\{y_1, ..., y_m\}$ be two different bases for S. Then:

- $\{x_1, ..., x_k\}$ are linearly independent in S, $\{y_1, ..., y_m\}$ span S, so $k \leq m$.
- $\{y_1, ..., y_m\}$ are linearly independent in S, $\{x_1, ..., x_k\}$ span S, so $m \leq k$.

It follows that $k = m$. The number of vectors required to form a basis for a subspace S is called the *dimension* of S, written $\dim(S)$. The trivial subspace $\{0\}$ has no basis, so we define its dimension to be 0. The next section shows that every nontrivial subspace $S \subseteq \mathbb{F}^n$ has a basis, and $1 \leq \dim(S) \leq n$.

5.4. Growing a basis via Lemma 5.2

Let S be a nontrivial subspace of $\mathbb{F}^n$. We show that S has a basis. Also, that basis has at least one vector and at most n vectors. The proof is a procedure for constructing a basis and uses Lemma 5.2 with the following facts:

> Fact 1: If $\{x_1, ..., x_k\}$ are vectors in subspace S, then $Span\{x_1, ..., x_k\} \subseteq S$.
> Fact 2: No collection of linearly independent vectors in $\mathbb{F}^n$ has more than n vectors.
> [Fact 1 holds by Lemma 3.2; Fact 2 holds by the wide matrix theorem]

Consider the following procedure for growing a basis for S:

- Step 1: Choose any nonzero vector from S and call this x_1. Such a vector exists because S is nontrivial. Then $\{x_1\}$ is linearly independent in S and Fact 1 ensures

$$Span\{x_1\} \subseteq S$$

  If $Span\{x_1\} = S$ then $\{x_1\}$ is a basis for S and we are done. Else, go to step 2.

- Step 2: If we get to this step, we know $\{x_1\}$ is linearly independent and $Span\{x_1\}$ is a strict subset of S. Choose any vector in S that is not in $Span\{x_1\}$ and call this x_2. Lemma 5.2 ensures $\{x_1, x_2\}$ are linearly independent. Fact 1 ensures

$$Span\{x_1, x_2\} \subseteq S$$

  If $Span\{x_1, x_2\} = S$ then $\{x_1, x_2\}$ is a basis for S and we are done. Else, go to step 3.

- Step $k \geq 3$: If we get to this step, we know $\{x_1, ..., x_{k-1}\}$ are linearly independent and $Span\{x_1, ..., x_{k-1}\}$ is a strict subset of S. Choose any vector in S that is not in $Span\{x_1, ..., x_{k-1}\}$ and call this x_k. Lemma 5.2 ensures $\{x_1, ..., x_k\}$ are linearly independent. Fact 1 ensures

$$Span\{x_1, ..., x_k\} \subseteq S$$

  If $Span\{x_1, ..., x_k\} = S$ then $\{x_1, ..., x_k\}$ is a basis for S and we are done. Else, go to the next step.

This procedure of growing a basis must stop on or before step n because Fact 2 ensures we cannot have more than n linearly independent vectors in $\mathbb{F}^n$. Thus, $1 \leq \dim(S) \leq n$.

5.5. Shrinking down to a basis via Lemma 5.1

Suppose we have a nontrivial subspace S and vectors $\{y_1, ..., y_m\}$ such that

$$Span\{y_1, ..., y_m\} = S$$

However, suppose these vectors are not linearly independent (so they are not a basis for S). Lemma 5.1 ensures that either $y_1 = 0$, or there is a $j \in \{2, ..., m\}$ such that y_j is a linear combination of $y_1, ..., y_{j-1}$. If $y_1 = 0$ we can throw y_1 away without affecting the span of the remaining vectors. On the other hand, if $y_1 \neq 0$ then there is a y_j as stated above, and we can remove y_j without affecting the span. In either case, we reduce the number of vectors by 1 while the span of the remaining vectors is still S. We keep going until one of two things occurs:

- The remaining vectors are linearly independent. The result is a basis for S.
- There is only one remaining vector. That vector cannot be 0 because its span is the original nontrivial subspace S. So that one nonzero vector is a linearly independent collection that spans S. The result is a basis for S.

Useful consequence: Let A be a $m \times n$ matrix with entries in some field. The n columns of A span $Col(A)$. If A is the all-zero matrix then $Col(A) = \{0\}$ and $\dim Col(A) = 0$. If A is not the all-zero matrix we do the following: Starting with the n columns of A, we repeatedly throw away any column that is a linear combination of the others to shrink down to a basis for $Col(A)$. Let r be the number of vectors in this basis. Then $r \leq n$.

5.6. Using dimension to establish that vectors form a basis

A useful consequence of Lemma 5.2 is that if $\{x_1, ..., x_k\}$ are linearly independent vectors in some subspace S, and if $y \in S$ but $y \notin Span\{x_1, ..., x_k\}$, then $\{x_1, ..., x_k, y\}$ are linearly independent in S (this is proven from first principles in Exercise 5.51).

Suppose we have a subspace S and we know $\dim(S) = k$ for some positive integer k. Suppose we also have linearly independent vectors $\{x_1, ..., x_k\}$ in S. Then they must form a basis for S. [Why? The reason is contained in the previous paragraph.] For example, $\{(1, 1), (1, 2)\}$ are linearly independent in $\mathbb{R}^2$, so they must be a basis for $\mathbb{R}^2$. Likewise, any n linearly independent vectors in $\mathbb{F}^n$ must be a basis for $\mathbb{F}^n$.

LEMMA 5.3. *Fix $k \in \mathbb{N}$. Let S be a subspace with $\dim(S) = k$.*

a) If $\{x_1, ..., x_k\}$ are linearly independent vectors in S, then they are also a basis for S.

b) If $\{y_1, ..., y_k\}$ are vectors that span S, then they are also a basis for S.

PROOF. This is left as Exercise 5.52. □

5.7. The rank of a matrix

Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$.

DEFINITION 5.1. *The* rank *of A, written $rank(A)$, is defined as the dimension of $Col(A)$.*

Note that $\text{rank}(A) = 0$ if and only if A is an all-zero matrix. If A is not an all-zero matrix then its rank is a positive integer.

LEMMA 5.4. *Let A be a $m \times n$ matrix with entries in some field. Let B be another matrix with size such that AB makes sense. Then*

a) $rank(A) \leq \min\{m, n\}$

b) $rank(A) = rank(A^\top)$

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

Lemma 5.4b means the number of linearly independent columns we can find in A is equal to the number of linearly independent rows we can find in A. Lemma 5.4c can be used with induction to prove that if $A_1, A_2, ..., A_k$ are matrices with sizes such that the product $A_1A_2 \cdots A_k$ makes sense, then

$$rank(A_1A_2 \cdots A_k) \leq \min\{rank(A_1), rank(A_2), ..., rank(A_k)\}$$

We prove Lemma 5.4 through a sequence of six claims.

Claim 1: $\text{rank}(A) \leq \min\{m, n\}$.

PROOF. Since $Col(A) \subseteq \mathbb{F}^m$, we know $dim(Col(A)) \leq m$ and so $\text{rank}(A) \leq m$. We now show $\text{rank}(A) \leq n$. If A is an all-zero matrix then $\text{rank}(A) = 0$ and the result is trivial. Else, the n columns of A can be reduced to a basis for $Col(A)$, so the number of vectors in this basis for $Col(A)$ is less than or equal to n. □

Claim 2: $Col(AB) \subseteq Col(A)$.

PROOF. Fix $y \in Col(AB)$. Then there is an x such that $y = ABx$. Then

$$y = ABx = A(Bx) \in Col(A)$$

and so $y \in Col(A)$. □

Claim 3: $\text{rank}(AB) \leq \text{rank}(A)$.

PROOF. This follows immediately from Claim 2. □

Claim 4: $\text{rank}(A^\top) \leq \text{rank}(A)$.

PROOF. If A is an all-zero matrix then $\text{rank}(A) = 0$, $\text{rank}(A^\top) = 0$, and the result is trivial. Now suppose A is not an all-zero matrix. Define $r = \text{rank}(A)$. Let U be a $m \times r$ matrix with columns that form a basis for $Col(A)$. Then every column of A is a linear combination of the columns of U, so

$$A = UC$$

for some $r \times n$ matrix C. Taking transposes gives

$$A^\top = C^\top U^\top$$

and so

$$\text{rank}(A^\top) = \text{rank}(C^\top U^\top) \overset{(a)}{\leq} rank(C^\top) \overset{(b)}{\leq} \min\{n, r\} \leq r$$

where inequality (a) uses Claim 3; inequality (b) uses Claim 1 and the fact that $C^\top$ has size $n \times r$. □

Claim 5: $\text{rank}(A^\top) = \text{rank}(A)$.

PROOF. This follows by applying Claim 4 to $A^\top$. □

Claim 6: $\text{rank}(AB) \leq \text{rank}(B)$.

PROOF. Using Claim 5 we have

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

where the inequality uses Claim 3. □

An alternative proof that $\text{rank}(AB) \leq \text{rank}(B)$ that does not use the transpose is in Exercise 5.42.

5.7.1. Rank-nullity theorem for matrices

THEOREM 5.3. *(rank-nullity for matrices) Let A be a $m \times n$ matrix with entries in $\mathbb{F}$. Let $r = rank(A)$ and $k = \dim Null(A)$.*

a) If $x_1, ..., x_r, z_1, ..., z_k$ are vectors in $\mathbb{F}^n$ and $y_1, ..., y_r$ are vectors in $\mathbb{F}^m$ such that

- *$\{z_1, ..., z_k\}$ is a basis for $Null(A)$.*

- $\{y_1, ..., y_r\}$ *is a basis for* $Col(A)$.
- $\{x_1, ..., x_r\}$ *satisfy* $Ax_i = y_i$ *for all* $i \in \{1, ..., r\}$

then $\{x_1, ..., x_r, z_1, ..., z_k\}$ *is a basis for* $\mathbb{F}^n$.

b) We have

$$\boxed{rank(A) + \dim Null(A) = n}$$

PROOF. To prove part (a), we first show $\{x_1, ..., x_r, z_1, ..., z_k\}$ are linearly independent. Suppose there are scalars $a_1, ..., a_r$ and $b_1, ..., b_k$ such that

$$\sum_{i=1}^{r} a_i x_i + \sum_{j=1}^{k} b_j z_j = 0 \quad \text{(Eq. *)}$$

We want to show $a_i = 0$ and $b_j = 0$ for all i, j. Multiplying the above equation by A gives

$$\sum_{i=1}^{r} a_i \underbrace{Ax_i}_{y_i} + \sum_{j=1}^{k} b_j \underbrace{Az_j}_{0} = 0$$

Substituting $Ax_i = y_i$ and $Az_j = 0$ (since the z_j vectors are in $Null(A)$) gives

$$\sum_{i=1}^{r} a_i y_i = 0$$

which immediately implies $a_i = 0$ for all i because we are told that $\{y_1, ..., y_r\}$ are linearly independent. Substituting $a_i = 0$ into (Eq. *) gives

$$\sum_{j=1}^{k} b_j z_j = 0$$

which immediately implies $b_j = 0$ for all j because we are told that $\{z_1, ..., z_k\}$ are linearly independent.

To complete the proof of part (a), it remains to show $\{x_1, ..., x_r, z_1, ..., z_k\}$ span $\mathbb{F}^n$. Fix $v \in \mathbb{F}^n$. We show v is a linear combination of $\{x_1, ..., x_r, z_1, ..., z_k\}$. Since $Av \in Col(A)$ and $\{y_1, ..., y_r\}$ span $Col(A)$ we have

$$Av = \sum_{i=1}^{r} a_i y_i$$

for some scalars $a_1, ..., a_r$. Define vector w by

$$w = \sum_{i=1}^{r} a_i x_i$$

Then

$$Aw = \sum_{i=1}^{r} a_i A x_i = \sum_{i=1}^{r} a_i y_i = Av$$

Therefore, v and w are two vectors in $\mathbb{F}^n$ that satisfy $Av = Aw$. Then $v - w \in Null(A)$ and so, because $\{z_1, ..., z_k\}$ span $Null(A)$, we have

$$v - w = \sum_{j=1}^{k} b_j z_j$$

for some scalars $b_1, ..., b_k$. So

$$v = w + \sum_{j=1}^{k} b_j z_j = \sum_{i=1}^{r} a_i x_i + \sum_{j=1}^{k} b_j z_j$$

where we have substituted the definition of w. So v is a linear combination of $\{x_1, ..., x_r, z_1, ..., z_k\}$.

To prove part (b), assume $r \geq 1$ and $k \geq 1$. Then there are vectors $\{x_1, ..., x_r, z_1, ..., z_k\}$ that satisfy the requirements of part (a) [why?]. So they form a basis for $\mathbb{F}^n$, so $r + k = n$. The cases $r = 0$ and $k = 0$ are left as exercises. □

For example, in the binary field $\mathbb{F}_2 = \{0, 1\}$, if a 10×20 matrix A with entries in $\mathbb{F}_2$ has $\dim Null(A) = 12$, then the rank-nullity theorem implies

$$\text{rank}(A) = 20 - 12 = 8$$

and so there are 2^8 vectors in $Col(A)$.

For any 10×20 matrix A (in any field $\mathbb{F}$), we know $\text{rank}(A) \leq 10$ (because $Col(A) \subseteq \mathbb{F}^{10}$), so the rank-nullity theorem implies $\dim Null(A) = 20 - rank(A) \geq 10$.

5.7.2. Terminology

Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$.

- We say A has *full row rank* if $\text{rank}(A) = m$ = Number of rows.
- We say A has *full column rank* if $\text{rank}(A) = n$ = Number of columns.
- We say A has *full rank* if $\text{rank}(A) = \min\{m, n\}$.

5.8. Equivalent conditions for invertibility

The following theorem is presented in the style of Lay, Lay, and McDonald (2016).

THEOREM 5.4. *Let A be a square $n \times n$ matrix with entries in some field $\mathbb{F}$. The following are equivalent:*

1. *A is invertible*
2. *There is a square matrix B such that $BA = I$*
3. *$Null(A) = \{0\}$*
4. *A has linearly independent columns*
5. *$Col(A) = \mathbb{F}^n$*
6. *$Ax = b$ has a solution for all $b \in \mathbb{F}^n$*
7. *There is a square matrix C such that $AC = I$*
8. *There is a square matrix D such that $DA^\top = I$*
9. *$A^\top$ is invertible*
10. *A has linearly independent rows*
11. *$rank(A) = n$*

PROOF. We first show that

$$1 \implies 2 \implies 3 \implies 4 \implies 5 \implies 6 \implies 7 \implies 8 \implies 9 \implies 1$$

- $1 \implies 2$: This follows from the definition of invertible.
- $2 \implies 3$: Fix $x \in Null(A)$. Then $Ax = 0$. Multiplying by B gives $BAx = 0$. But we are told $BA = I$ and so $Ix = 0$, that is, $x = 0$. So 0 is the only vector in $Null(A)$.
- $3 \implies 4$: Since $Null(A) = \{0\}$, Lemma 3.7 ensures the columns of A are linearly independent.
- $4 \implies 5$: The columns of A are n linearly independent vectors in $\mathbb{F}^n$, so they must be a basis for $\mathbb{F}^n$ (by Lemma 5.3).
- $5 \implies 6$: Since $Col(A) = \mathbb{F}^n$, every $b \in \mathbb{F}^n$ is in $Col(A)$, meaning there is an x such that $b = Ax$.
- $6 \implies 7$: Since $Ax = b$ has a solution for all b, for each standard basis vector e_i we have a vector x_i such that $Ax_i = e_i$. So $Ax_1 = e_1$, $Ax_2 = e_2$, ..., $Ax_n = e_n$. The $n \times n$ identity matrix I has columns e_i:

$$I = \begin{bmatrix} e_1 & e_2 & \dots & e_n \end{bmatrix}$$

Define C as a matrix with columns x_i:

$$C = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$$

Then

$$AC = \begin{bmatrix} Ax_1 & Ax_2 & \dots & Ax_n \end{bmatrix} = I$$

- 7 $\implies$ 8: Taking the transpose of $AC = I$ yields $C^\top A^\top = I$. Define $D = C^\top$.
- 8 $\implies$ 9: We know there is a square matrix D such that $DA^\top = I$. Therefore, the condition 2 holds for the matrix $A^\top$. We have already proven that 2 $\implies$ 7. So there is a square matrix C such that $A^\top C = I$. Lemma 2.3 proves that $D = C = (A^\top)^{-1}$.
- 9 $\implies$ 1: We already know (by Lemma 2.7) that A is invertible if and only if $A^\top$ is invertible.

Thus, statements 1, 2, 3, 4, 5, 6, 7, 8, 9 are equivalent. As an exercise, the reader can use this to prove 10 $\iff$ 9 (note that A has linearly independent rows if and only if $A^\top$ has linearly independent columns). Finally, the rank-nullity theorem can be used to prove 11 $\iff$ 3. [Alternatively, one can use Lemma 5.2 to prove 11 $\iff$ 5.] □

5.9. Random matrices

Let $\mathbb{F}_q$ be a finite field of size q. [A remarkable theorem due to mathematicians Galois and Moore is that a finite field of size q exists if and only if $q = p^k$ for some prime p and some $k \in \mathbb{N}$.] Let A be a *random* $m \times n$ matrix with entries that are chosen independently and uniformly over the scalars in $\mathbb{F}_q$. Then $A = (A_{ij})$ where $\{A_{ij}\}$ are random scalars that are independent and identically distributed (i.i.d.) with

$$P[A_{ij} = a] = \frac{1}{q} \quad \forall a \in \mathbb{F}_q$$

where $P[E]$ denotes the probability of an event E in the sample space of our probability experiment.

This probability law is equivalent to randomly choosing the entire matrix A equally likely over the q^{mn} matrices of size $m \times n$ with entries in $\mathbb{F}_q$. Equivalently, we can independently choose each column of A equally likely over the q^m possible vectors in $\mathbb{F}_q^m$.

5.9.1. Probability of linearly independent columns

Using field $\mathbb{F}_q$, what is the probability that the randomly chosen $m \times n$ matrix A has linearly independent columns? If $n > m$, the wide matrix theorem tells us the probability is zero. So let us assume $n \leq m$ (the number of columns is less than or equal to the number of

rows).

Write the random matrix A in terms of its columns:

$$A = \begin{bmatrix} | & | & \cdots & | \\ C_1 & C_2 & \cdots & C_n \\ | & | & \cdots & | \end{bmatrix}$$

For each $i \in \{1, \ldots, n\}$ we have

$$P[C_i = x] = \frac{1}{q^m} \quad \forall x \in \mathbb{F}_q^m$$

Define the following events:

$$\begin{aligned} L &= \{A \text{ has linearly independent columns}\} \\ E_1 &= \{C_1 \neq 0\} \\ E_2 &= \{C_2 \notin Span\{C_1\}\} \\ E_3 &= \{C_3 \notin Span\{C_1, C_2\}\} \\ &\vdots \\ E_n &= \{C_n \notin Span\{C_1, C_2, \ldots, C_{n-1}\}\} \end{aligned}$$

Lemma 5.2 implies

$$L = \cap_{i=1}^n E_i$$

So by the rules of conditional probability:

$$P[L] = P[E_1]P[E_2|E_1]P[E_3|E_1 \cap E_2]\cdots P[E_n| \cap_{i=1}^{n-1} E_i] \tag{5.2}$$

Since C_1 is chosen equally likely over all q^m vectors in $\mathbb{F}_q^m$, and 0 is one particular vector in $\mathbb{F}_q^m$, we have

$$P[E_1] = P[C_1 \neq 0] = \frac{q^m - 1}{q^m} \tag{5.3}$$

The events $\{E_1, \ldots, E_n\}$ are *dependent*. For example, whether or not E_1 is true determines the size of $Span\{C_1\}$, which affects the likelihood of event E_2. Indeed, if we are told E_1 is true, then $C_1 \neq 0$ and there are exactly q vectors in $Span\{C_1\}$ (one for each scalar multiple of C_1). On the other hand, if E_1 is false (meaning its complement E_1^c is true) then $C_1 = 0$ and $Span\{C_1\}$ consists of the single vector 0.

$$E_1 \implies (C_1 \neq 0) \implies |Span\{C_1\}| = q \tag{5.4}$$

$$E_1^c \implies (C_1 = 0) \implies |Span\{C_1\}| = 1 \tag{5.5}$$

where $|S|$ denotes the number of elements in a set S. Thus

$$P[E_2|E_1] = \frac{q^m - q}{q^m} \tag{5.6}$$

$$P[E_2|E_1^c] = \frac{q^m - 1}{q^m} \tag{5.7}$$

where the numerator $q^m - q$ in (5.6) is the number of vectors in $\mathbb{F}_q^m$ that are *not* in $Span\{C_1\}$ (given E_1 is true); the numerator $q^m - 1$ in (5.7) is the number of vectors in $\mathbb{F}_q^m$ that are *not* in $Span\{C_1\}$ (given E_1^c is true).

Similarly, if $E_1 \cap E_2$ is true then $\{C_1, C_2\}$ are linearly independent and $Span\{C_1, C_2\}$ is a subspace of $\mathbb{F}_q^m$ of dimension 2, so (by Lemma 3.8):

$$(E_1 \cap E_2) \implies \left(|Span\{C_1, C_2\}| = q^2\right) \implies P[E_3|E_1 \cap E_2] = \frac{q^m - q^2}{q^m}$$

Generally, for any $d \in \{1, ..., n-1\}$, if $\cap_{i=1}^d E_i$ is true then $\{C_1, ..., C_d\}$ are linearly independent and $Span\{C_1, ..., C_d\}$ is a subspace of dimension d, so (by Lemma 3.8):

$$\left(\cap_{i=1}^d E_i\right) \implies |Span\{C_1, C_2, ..., C_d\}| = q^d$$

Thus

$$P[E_{d+1}|\cap_{i=1}^d E_i] = \frac{q^m - q^d}{q^m} \quad \forall d \in \{1, ..., n-1\} \tag{5.8}$$

Substituting (5.8),(5.3) into (5.2) gives

$$\begin{aligned} P[L] &= \left(\frac{q^m - 1}{q^m}\right)\left(\frac{q^m - q}{q^m}\right)\left(\frac{q^m - q^2}{q^m}\right)\cdots\left(\frac{q^m - q^{n-1}}{q^m}\right) \\ &= \left(1 - \frac{1}{q^m}\right)\left(1 - \frac{1}{q^{m-1}}\right)\left(1 - \frac{1}{q^{m-2}}\right)\cdots\left(1 - \frac{1}{q^{m-n+1}}\right) \end{aligned} \tag{5.9}$$

5.9.2. Probability that A is invertible

In the special case when A is a square $n \times n$ matrix, the columns of A are linearly independent if and only if A is invertible. Therefore, the probability that A is invertible can be found by substituting $m = n$ into formula (5.9):

$$P[A \text{ is invertible}] = \left(1 - \frac{1}{q}\right)\left(1 - \frac{1}{q^2}\right)\left(1 - \frac{1}{q^3}\right)\cdots\left(1 - \frac{1}{q^n}\right)$$

[For convenience, factors are reversed here in comparison to (5.9).] This is summarized in the following theorem.

THEOREM 5.5. *Consider a finite field $\mathbb{F}_q$ (where $q = p^k$ for some prime p and some $k \in \mathbb{N}$). Fix $n \in \mathbb{N}$. If $A = (A_{ij})$ is a random $n \times n$ matrix, where entries are chosen independently and uniformly over all q scalars in $\mathbb{F}_q$, then*

$$\boxed{P[A \text{ is invertible}] = \prod_{i=1}^{n} \left(1 - \frac{1}{q^i}\right)}$$

Table 5.1 gives numerical values for various q and n. As a rule of thumb, when $q \geq 16$ the result does not significantly depend on n and

$$P[A \text{ is invertible}] \approx 1 - \frac{1}{q} \quad \forall q \geq 16, \forall n \geq 3$$

	n=3		n=4		n=100
q	P[invertible]	q	P[invertible]	q	P[invertible]
2	0.3281	2	0.3076	2	0.2888
4	0.6921	4	0.6894	4	0.6885
8	0.8596	8	0.8594	8	0.8594
16	0.9336	16	0.9336	16	0.9336
32	0.9678	32	0.9678	32	0.9678
64	0.9841	64	0.9841	64	0.9841
128	0.9921	128	0.9921	128	0.9921
256	0.9961	256	0.9961	256	0.9961

TABLE 5.1. Probability a random $n \times n$ matrix is invertible. Entries are in $\mathbb{F}_q$.

5.9.3. Random matrices with the field $\mathbb{R}$

Now consider choosing a random $m \times n$ matrix with entries in the field $\mathbb{R}$. Assume $n \leq m$ so the matrix is either square or tall. Specifically, suppose $A = (A_{ij})$ is a random $m \times n$ matrix with entries A_{ij} that are i.i.d. and chosen according to some *continuous distribution*, so for each (i, j) we have

$$P[A_{ij} \leq x] = F(x) \quad \forall x \in \mathbb{R}$$

where $F(x)$ is some continuous cumulative distribution function (CDF), such as the CDF for a Gaussian random variable, or an exponential random variable, or a random variable

that is uniformly distributed over some interval $[a, b]$ where $a, b \in \mathbb{R}$ and $a < b$. Then, since A is either square or tall, we have

$$P[A \text{ has linearly independent columns}] = 1 \tag{5.10}$$

This is because, by (5.2),

$$\begin{aligned} &P[A \text{ has linearly independent columns}] \\ &= P[E_1]P[E_2|E_1]P[E_3|E_1 \cap E_2]\cdots P[E_n| \cap_{i=1}^{n-1} E_i] \end{aligned} \tag{5.11}$$

It can be shown that each factor in the right-hand-side of the above equality is 1. Specifically, let $C_1, ..., C_n$ be the columns of A. To show that $P[E_1] = 1$, it suffices to show $P[E_1^c] = 0$. But $P[E_1^c] = P[C_1 = 0]$. Also, $C_1 = (A_{11}, A_{21}, ..., A_{m1}) \in \mathbb{R}^m$ and

$$P[C_1 = 0] = P[A_{11} = 0]P[A_{21} = 0]\cdots P[A_{m1} = 0] = 0$$

where we recall that, because each A_{i1} is a random variable with a continuous CDF, we have $P[A_{i1} = x] = 0$ for all $x \in \mathbb{R}$ and all $i \in \{1, ..., m\}$.

We omit a formal proof that the remaining factors in the right-hand-side of (5.11) are also 1. However, informally, it holds because integrating a probability density function (PDF) over any zero-volume subset of $\mathbb{R}^m$ yields zero, and

- $P[E_2|E_1] = 1$ since if E_1 is true then $Span\{C_1\}$ is a 1-dimensional line within $\mathbb{R}^m$. This 1-dimensional line has zero volume and hence zero probability within the m-dimensional space $\mathbb{R}^m$ (since $1 < m$).
- $P[E_3|E_1 \cap E_2] = 1$ since if $E_1 \cap E_2$ is true then $Span\{C_1, C_2\}$ is a 2-dimensional subspace, which has zero volume and hence zero probability within the m-dimensional space $\mathbb{R}^m$ (since $2 < m$).

  ⋮
- $P[E_n| \cap_{i=1}^{n-1} E_i] = 1$ since if $\cap_{i=1}^{n-1} E_i$ is true then $Span\{C_1, ..., C_{n-1}\}$ is a $(n-1)$-dimensional subspace, which has zero volume and hence zero probability within the m-dimensional space $\mathbb{R}^m$ (since $n - 1 < m$).

[Equation (5.10) holds for any continuous CDF, regardless of existence of a PDF. A continuous CDF does not necessarily have a PDF. However, most do, and integrating the PDF over a zero-volume subset of $\mathbb{R}^m$ provides the best intuition.]

5.10. The chain of null spaces for powers of square matrices

Fix $\mathbb{F}$ as a field. Let A be a $n \times n$ matrix with entries in $\mathbb{F}$. For a given positive integer k, let A^k denote the matrix multiplication of A with itself k times over:

$$A^2 = AA$$
$$A^3 = AAA = A^2A$$
$$A^4 = AAAA = A^3A$$

and so on. Since A is a square matrix, A^k is also a square matrix with the same size as A. The notation A^k is only defined when A is square (else, the multiplication of A with itself does not make sense).

It is easy to observe that for all positive integers k:

$$\{0\} \subseteq Null(A^k) \subseteq Null(A^{k+1}) \subseteq \mathbb{F}^n \tag{5.12}$$

Indeed, if $x \in Null(A^k)$ then

$$A^k x = 0$$

and multiplying the above equation by A gives

$$A^{k+1}x = 0$$

so $x \in Null(A^{k+1})$, proving (5.12). From (5.12) it follows that

$$\{0\} \subseteq Null(A) \subseteq Null(A^2) \subseteq Null(A^3) \subseteq Null(A^4) \cdots \subseteq \mathbb{F}^n \tag{5.13}$$

Thus, the null space either stays the same or strictly increases as we take more powers of A.

LEMMA 5.5. *Fix n as a positive integer and let A be a $n \times n$ matrix with entries in some field.*

a) If there is a positive integer m such that $Null(A^{m+1}) = Null(A^m)$ then

$$Null(A^k) = Null(A^m) \quad \forall k \geq m$$

b) If $Null(A) = \{0\}$ then

$$Null(A^k) = \{0\} \quad \forall k \geq 1$$

c) We have

$$Null(A^k) = Null(A^n) \quad \forall k \geq n$$

PROOF. The proof of (a) is similar to Exercise 5.29. To prove (b), suppose $Null(A) = \{0\}$. We first show that $Null(A^2) = \{0\}$. Suppose x is a vector in $Null(A^2)$ (we show x must be 0). Since $x \in Null(A^2)$ we know $A^2x = 0$, and so

$$A \cdot Ax = 0$$

which means

$$Ax \in Null(A) \tag{5.14}$$

But $Null(A) = \{0\}$, so (5.14) implies $Ax = 0$. Thus, $x \in Null(A) = \{0\}$, so $x = 0$. This completes the proof that $Null(A^2) = \{0\}$. By part (a), we must have $Null(A^k) = \{0\}$ for all $k \geq 1$, which completes the proof of (b).

The proof of (c) is left as an exercise. Hint: consider the case $\dim Null(A) = 0$ and the opposite case $\dim Null(A) \geq 1$, and use reasoning similar to the discussion paragraph given below. □

The lemma implies that, in the chain of subspace inclusions (5.13), there is a finite (possibly empty) sequence of steps where the null space *strictly increases* (which means the dimension increases by at least one at each step) until the first step when the null space stays the same. The number of strictly increasing steps is finite because the dimension increases by at least one at every such step, while the largest possible dimension is n (since $Null(A^k) \subseteq \mathbb{F}^n$ for all k). Once we reach a step where the null space remains the same as the previous step, it stays the same at each step forever after. We say the null space *stabilizes* at that point, and call the stabilized null space the *limiting null space*. Since the dimension increases by at least one at each strictly increasing step, the limiting null space must be reached on or before step n. So the limiting null space is $Null(A^n)$.

For example, the identity matrix has $I^k = I$ for all k, so if $A = I$ the chain of inclusions (5.13) reduces to

$$\{0\} = Null(I) = Null(I) = Null(I) = \ldots$$

For another example, consider the 3×3 matrix A defined by

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix} \quad , \quad A^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \quad , \quad A^k = A^2 \quad \forall k \geq 2$$

For this matrix A we have

$$\{0\} \subsetneq Null(A) \subsetneq Null(A^2) = Null(A^3) = Null(A^4) = \ldots$$

where $C \subsetneq D$ denotes C is a *strict subset* of D (meaning every element of C is also in D, but there is at least one element of D that is not in C). Indeed, it is not difficult to show

$$Null(A) = Span\{(0,1,0)\}$$
$$Null(A^2) = Span\{(0,1,0),(0,0,1)\}$$

As an exercise, the reader is invited to find a specific square matrix A where the null space dimension increases by *more than 1* at a given step, such as a matrix A that satisfies $\dim Null(A^2) = 2 + \dim Null(A)$ (see Exercise 5.31).

5.11. Exercises

EXERCISE 5.1. *Let A be a $n \times n$ matrix with entries in some field. Show that* $\dim Null(A^2) \geq \dim Null(A)$*, where $A^2 = AA$.*

EXERCISE 5.2. *Give an example of a nonsquare matrix A with at least two rows and at least two columns such that the equation $Ax = b$ has a solution for all vectors b.*

EXERCISE 5.3. *Give an example matrix A (using the field $\mathbb{R}$) such that* $\dim Null(A) = 5$ *and* $\dim Col(A) = 2$.

EXERCISE 5.4. *(Example matrix in $\mathbb{F}_3$) Let $\mathbb{F}_3 = \{0,1,2\}$ with mod 3 arithmetic.*

a) Give an example matrix A and column vector b that satisfy all of the following:

- *A is a 3×3 matrix with entries in $\mathbb{F}_3$.*
- *All entries of A are nonzero.*
- *b is a 3×1 column vector with entries in $\mathbb{F}_3$.*
- *The equation $Ax = b$ has no solutions in mod 3 arithmetic.*

b) How many vectors are in $Null(A)$? How many in $Col(A)$?

EXERCISE 5.5. *(Example matrix) Give an example of a matrix A (with some field $\mathbb{F}$) that meets the following criteria:*

- *$Ax = b$ has exactly 27 solutions for some vectors b, and no solutions for other vectors b.*
- *$Col(A)$ has dimension 2.*
- *All columns of A are different.*

EXERCISE 5.6. *Suppose A and B are square matrices of the same size (they have elements in some field $\mathbb{F}$). Show that AB is invertible if and only if both A and B are invertible.*

EXERCISE 5.7. *Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Let G be a $m \times m$ invertible matrix with entries in $\mathbb{F}$.*

a) Show that $Null(A) = Null(GA)$.

b) Show that $rank(A) = rank(GA)$.

EXERCISE 5.8. *Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Let G be a $n \times n$ invertible matrix with entries in $\mathbb{F}$. Show that $rank(AG) = rank(A)$.*

EXERCISE 5.9. *Let A be a $m \times n$ matrix. Suppose W is an invertible $m \times m$ matrix.*

a) Show that A has linearly independent columns if and only if WA has linearly independent columns.

b) Explain why A cannot have linearly independent columns if it is a wide matrix.

c) Recall that A has linearly independent columns if and only if the first column is not the all-zero vector, the second column is not a multiple of the first, the third column is not a linear combination of the first two, and so on. Suppose the entries of A come from measured data. Explain why we expect this to be true when A is very tall.

EXERCISE 5.10. *Suppose A is a $m \times 5$ matrix with real entries and $b \in \mathbb{R}^m$. We are told two pieces of information:*

- $$Null(A) = Span\left\{\begin{bmatrix} 2 \\ 0 \\ -1 \\ 0 \\ 4 \end{bmatrix}, \begin{bmatrix} 7 \\ 7 \\ 1 \\ 9 \\ 0 \end{bmatrix}, \begin{bmatrix} 2 \\ 1 \\ 5 \\ 0 \\ 8 \end{bmatrix}\right\}$$
- *A particular solution to $Ax = b$ is the vector*

$$x^* = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 0 \end{bmatrix}$$

Find all solutions to $Ax = b$.

EXERCISE 5.11. *Give an example of a field $\mathbb{F}$ and a matrix A that simultaneously satisfies all of the following properties:*

- *Null*(A) *contains at least 7 vectors.*
- *Col*(A) *contains exactly 125 vectors.*
- *A has 4 rows, and no entries of A are* 0 *or* 1.

EXERCISE 5.12. *Using field* $\mathbb{F}_2 = \{0, 1\}$*, consider the* 4×5 *matrix:*

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

a) Find a last column that consists of a single 1 and that ensures Null(A) *has dimension 2.*

b) Find all last columns that ensure Null(A) *has dimension 1.*

EXERCISE 5.13. *Suppose A, B, C are possibly nonsquare matrices that satisfy C = BA. Suppose C has linearly independent columns.*

a) Show the columns of A must be linearly independent.

b) Give a specific example to show the columns of B need not be linearly independent.

EXERCISE 5.14. *(If Null*$(A) = \{0\}$*) Let A be a square matrix with entries in some field. Recall that* $A^2 = AA$*,* $A^3 = AAA$*, and so on. Suppose Null*$(A) = \{0\}$*.*

a) Show that Null$(A^2) = \{0\}$*.*

b) Show that Null$(A^k) = \{0\}$ *for all* $k \in \{1, 2, 3, \ldots\}$*.*

c) Give an example of a 2×2 *matrix A (and a specific field) such that Null*$(A) = \{0\}$*.*

d) Give an example of a 2×2 *matrix B (and a specific field) such that Null*$(B) \neq \{0\}$ *and Null*$(B^2) \neq$ *Null*(B)*.*

EXERCISE 5.15. *(If rank*$(A) =$ *rank*(A^2)*) Let A be a square matrix with entries in some field. Suppose rank*$(A) =$ *rank*(A^2)*. Prove that Null*$(A) =$ *Null*(A^2)*.*

EXERCISE 5.16. *Give an example of matrices A and B of different size that satisfy rank*$(A) =$ *rank*$(B) = 3$ *and* $\dim$ *Null*$(A) = \dim$ *Null*$(B) = 4$*.*

EXERCISE 5.17. *(Graph game) Consider a game played on the 5-node undirected graph shown in Fig. 5.1. The color of each node can be either orange or blue. Assume all 5 nodes start with color orange (as shown in the figure). Clicking on a node toggles the color of that node and its neighbors. For example, clicking on node 2 when the graph is in the all-orange state changes the color of nodes 2, 1, 3, 4 to blue, while node 5 remains orange. We want to find a sequence of clicks that makes all nodes blue, if possible.*

a) State the 5×5 adjacency matrix $A = (a_{ij})$*, where* $a_{ii} = 1$ *for all* $i \in \{1, \ldots, 5\}$*, and if* $i \neq j$ *we have* $a_{ij} = 1$ *if there is an edge between nodes i and j (and* $a_{ij} = 0$ *else). Argue that*

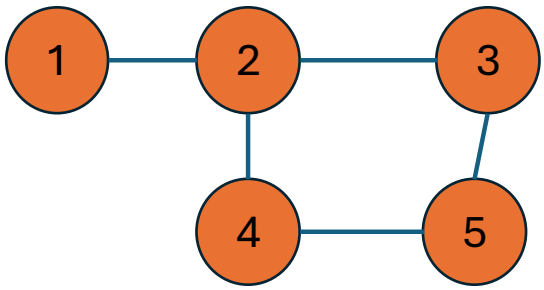


FIGURE 5.1. The 5-node graph for Exercise 5.17.

if $x = (x_1, x_2, x_3, x_4, x_5)$ is a vector with nonnegative integer entries, where x_i represents the number of times we click on node i, then Ax is a vector of nonnegative integers that represents the number of times the color of each node changes (where arithmetic is in the real field). Argue that to reach any desired state of colors in the nodes, it suffices to click on each node at most once.

b) Now treat $x \in \mathbb{F}_2^5$ where $x_i = 1$ if we click on node i, and 0 else. Argue that Ax, computed in the binary field, is the color state of the graph after clicking each node according to the vector x.

c) Is the all-blue state reachable from the all-orange state? If so, what is the smallest number of clicks required to turn all nodes blue?

d) Note that since A is symmetric we have $Row(A) = Col(A)$. Find a basis for $Col(A)$ (using the binary field). Find a basis for $Null(A)$ (using the binary field). Of the 2^5 different color-states, how many are reachable from the all-orange state?

EXERCISE 5.18. *For the 5-node graph of Fig. 5.1, suppose we start in state*

$$(orange, orange, orange, orange, blue)$$

(so node 5 is initially blue).

a) What is the smallest number of clicks to get to the all-orange state?

b) What is the smallest number of clicks to get to the all-blue state?

EXERCISE 5.19. *(A 9-node graph game) Consider a graph game with the same rules as in Exercise 5.17, but with graph shown in Fig. 5.2. Each node can be either orange or blue, and clicking a node toggles its color and the color of its neighbors.*

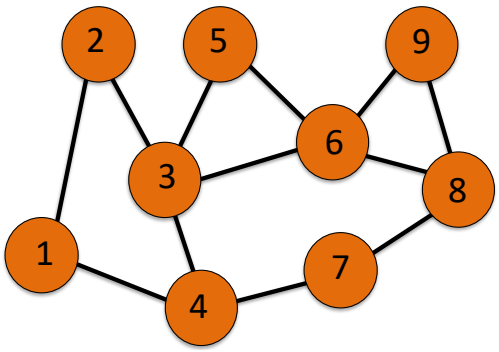


FIGURE 5.2. The 9-node graph for Exercise 5.19.

a) State the 9×9 adjacency matrix A. Remember that $a_{ii} = 1$ for all $i \in \{1, \ldots, 9\}$.

b) Find the shortest sequence of clicks that takes the 9-node graph from the all-orange state to the all-blue state.

c) From the all-orange state, how many of the 2^9 color-states are reachable?

EXERCISE 5.20. *Create a graph theory game problem similar to the previous exercises. Feel free to reuse one of the graphs, or make your own graph.*

EXERCISE 5.21. *(Erasure coding) Let $\mathbb{F}_2 = \{0, 1\}$ be the binary field. Let $x, y \in \mathbb{F}_2^n$ represent two files of binary data. Assume n is large (so each file is a significant portion of a hard drive). Define a new file $z \in \mathbb{F}_2^n$ by $z = x + y$ (using a mod-2 sum). The files x, y, z all have the same size. For reliable storage, we store x, y, z in three separate locations.*

a) Suppose the storage in one location is destroyed (so we lose either x, y, or z). Show we can recover both x and y from the remaining two files.

b) For each $i \in \{1, \ldots, n\}$, show that

$$\begin{bmatrix} x_i \\ y_i \\ z_i \end{bmatrix} = A \begin{bmatrix} x_i \\ y_i \end{bmatrix}, \quad A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 1 \end{bmatrix}$$

using arithmetic in the binary field. Argue that any two rows of A are linearly independent.

c) Suppose we lose file y. Argue that for each $i \in \{1, \ldots, n\}$ we have

$$\begin{bmatrix} x_i \\ z_i \end{bmatrix} = \tilde{A} \begin{bmatrix} x_i \\ y_i \end{bmatrix}$$

where $\tilde{A}$ is an invertible 2×2 matrix formed by removing a particular row of A. Thus, data recovery can be viewed as the matrix equation $[x_i; y_i] = \tilde{A}^{-1}[x_i; z_i]$ for each $i \in \{1, \ldots, n\}$.

d) Define $\mathbb{F}_3 = \{0, 1, 2\}$ as the size-3 field with mod-3 arithmetic. Construct a 4×2 matrix A with entries in $\mathbb{F}_3$ such that any two rows are linearly independent (you need to check all $\binom{4}{2}$ pairs of rows). Use this to show that any two vectors of trinary data $x, y \in \mathbb{F}_3^n$ can be encoded into four vectors $a, b, c, d \in \mathbb{F}_3^n$ such that the full data x, y can be recovered from any two of a, b, c, d. This is called erasure coding *because it can recover erased data as long as enough unerased data remains.*

EXERCISE 5.22. *(Can this matrix create an erasure code?) Consider the following matrix with entries in $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$:*

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

Using field $\mathbb{F}_5$, either find three linearly dependent rows, or prove that any three rows of A are linearly independent.

EXERCISE 5.23. *(Random erasure coding) This generalizes Exercise 5.21. Suppose we have some finite field $\mathbb{F}$. Fix positive integers m, k, n with $k < m$. Suppose we have k packets of data*

$x_1, \ldots, x_k \in \mathbb{F}^n$. *We want to encode this into* m *packets of* coded data $c_1, \ldots, c_m \in \mathbb{F}^n$ *such that all original data can be recovered from any* k *of the packets* $c_1, \ldots, c_m$. *Let* $x_i[j] \in \mathbb{F}$ *and* $c_i[j] \in \mathbb{F}$ *denote the* j*th component of packets* x_i *and* c_i, *respectively.*

a) Suppose A *is a* $m \times k$ *matrix with entries in* $\mathbb{F}$. *Suppose* A *has the property that any* k *rows are linearly independent. For each* $j \in \{1, \ldots, n\}$ *define*

$$\begin{bmatrix} c_1[j] \\ c_2[j] \\ \vdots \\ c_m[j] \end{bmatrix} = A \begin{bmatrix} x_1[j] \\ x_2[j] \\ \vdots \\ x_k[j] \end{bmatrix}$$

Argue that we can recover all original data from any k *of the coded packets* $c_1, \ldots, c_m$.

b) Suppose A *is the* $m \times k$ *matrix of part (a). Fix* $j \leq k$ *and let* $r_1, \ldots, r_j$ *be any particular distinct rows of* A. *We know that* $\{r_1, \ldots, r_j\}$ *are linearly independent. Define* $f = |\mathbb{F}|$ *as the positive integer field size. Find the number of vectors in* $Span\{r_1, \ldots, r_j\}$ *in terms of* f *and* j.

c) We now show that a matrix of the kind used in part (a) exists whenever the field size is large enough. Suppose we choose the coefficients of A randomly *by independently choosing each entry uniformly over all* f *elements of* $\mathbb{F}$. *Equivalently, each row is chosen independently and uniformly over all* f^k *choices of vectors in* $\mathbb{F}^k$. *Let* p *be the probability that the first* k *rows of* A *are linearly independent. Compute* p *in terms of* f *and* k. *Hint: Define* $B_1 = \{$*row 1 is not all zero*$\}$ *and for* $i \in \{2, \ldots, k\}$ *define* $B_i = \{$*row* i *is not in the span of rows* $1, \ldots, i-1\}$. *Recall that*

$$P[\cap_{i=1}^{k} B_i] = P[B_1]P[B_2|B_1]P[B_3|B_1 \cap B_2] \cdots P[B_k| \cap_{i=1}^{k-1} B_i]$$

Recall that $\cap_{i=1}^{j} B_i$ *is true if and only if rows* $1, \ldots, j$ *are linearly independent. Proceed to part (d) of this problem.*

d) Let p_f *denote the value of* p *in part (c) (treated as a function of* f*). By symmetry and the union bound, the probability that the randomly formed* $m \times k$ *matrix* A *contains a collection of* k *rows that is* not *linearly independent is less than or equal to* $\binom{m}{k}(1 - p_f)$. *Show that the limit of this as* $f \to \infty$ *is* 0. *Argue that if the field size* f *is large enough, there exists a matrix* A *of the kind needed in part (a).*[1]

EXERCISE 5.24. *(Bits-to-bits erasure coding with* $\mathbb{F}_4$*) This problem builds on Exercise 5.21. We have two blocks of binary data* $x = [x_1, \ldots, x_n]$ *and* $y = [y_1, \ldots, y_n]$. *Suppose* n *is even.*

a) Using the field $\mathbb{F}_4 = \{0, 1, a, b\}$ *with arithmetic defined by Exercise 1.17, design an erasure coding scheme to encode this into four blocks of binary data* c_1, c_2, c_3, c_4, *each of size* n, *such that recovery of any two of the blocks enables complete recovery of the original data* x *and* y. *Give*

[1] In practice, it is common to use the $k \times k$ identity matrix for the first k rows of A. The remaining $m - k$ rows can be formed by a random search.

your encoding for the case $n = 10$ *and*

$$x = [0011011010]$$
$$y = [1001111011]$$

You may want to use, for example

$$00 \leftrightarrow 0, \quad 01 \leftrightarrow 1, \quad 10 \leftrightarrow a, \quad 11 \leftrightarrow b$$

so $x = [0, b, 1, a, a]$ *and* $y = [a, 1, b, a, b]$. *This involves the design of a* 4×2 *matrix* A *with elements in* $\mathbb{F}_4$ *such that any two rows of* A *are linearly independent (together with carrying out the nonstandard arithmetic in* $\mathbb{F}_4$*).*[2] *To be compatible with part (b), the last two rows of your matrix* A *should not have any zeros.*

b) Design a decoding scheme. You may want to use the 2×2 *matrix inversion formula in Exercise 2.22. Explain how to decode to find blocks* x *and* y *in the case when blocks* c_1 *and* c_2 *have been destroyed, so we only have blocks* c_3 *and* c_4 *(where* c_3 *and* c_4 *are blocks different from* x, y*).*

EXERCISE 5.25. *(*$(n, 3)$ *erasure code) Fix* n *as an integer (assume* $n \geq 4$*). Let* q *be a prime number such that* $q \geq n$*. Using field* $\mathbb{F}_q = \{0, 1, \ldots, q-1\}$ *with mod* q *arithmetic, show that the following matrix, called a* Vandermonde matrix*, can be used for a* $(n, 3)$ *erasure code, meaning that any* 3 *rows are linearly independent:*

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 2^2 \\ 1 & 3 & 3^2 \\ \vdots & \vdots & \vdots \\ 1 & (n-1) & (n-1)^2 \end{bmatrix}$$

so each row $i \in \{1, \ldots, n\}$ *has the form* $(1, (i-1), (i-1)^2)$*, which is useful.*[3]

EXERCISE 5.26. *(Erasure code from Vandermonde matrix) Let* $\mathbb{F}$ *be a field. Let* A *be a* $n \times k$ *matrix with entries in* $\mathbb{F}$ *(assume* $n > k > 0$*). Suppose any* k *rows of* A *are linearly independent. Let* B *be an invertible* $k \times k$ *matrix (with entries in* $\mathbb{F}$*).*

a) Show that any k *rows of* AB *are linearly independent.*

[2] Note that such a matrix A is impossible in $\mathbb{F}_2$. Exercise 5.21 shows it is possible in $\mathbb{F}_3$, but that uses trinary data rather than bits-to-bits encoding.

[3] This fact about the $n \times 3$ Vandermonde matrix A holds more generally for any finite field $\mathbb{F}_q$ with $q \geq n$ (such as $q = 2^k \geq n$). In that case column 2 of A is replaced by $(s_1, s_2, \ldots, s_n)$ and column 3 is replaced by $(s_1^2, s_2^2, \ldots, s_n^2)$, where $\{s_1, \ldots, s_n\}$ are any distinct scalars in $\mathbb{F}_q$.

b) Consider the following 7×4 Vandermonde matrix *A, with entries in* $\mathbb{F}_7 = \{0,1,2,3,4,5,6\}$ *(with mod 7 arithmetic):*

$$A = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 2^2 & 2^3 \\ 1 & 3 & 3^2 & 3^3 \\ 1 & 4 & 4^2 & 4^3 \\ 1 & 5 & 5^2 & 5^3 \\ 1 & 6 & 6^2 & 6^3 \end{bmatrix}$$

It is known that any 4 *rows of A are linearly independent. Write A with all entries in* $\mathbb{F}_7$ *(for example,* $2^3 = 1$*). Use part (a) to convert to a new matrix* $\tilde{A}$ *with the following properties:*

- $\tilde{A}$ *has size* 7×4 *(with all entries in* $\mathbb{F}_7$*).*
- *Any 4 rows of* $\tilde{A}$ *are linearly independent.*
- *The first 4 rows of* $\tilde{A}$ *form the* 4×4 *identity matrix.*[4]

EXERCISE 5.27. *(Implementation of* $(5,3)$ *erasure code with binary-to-*5*-ary conversion) Consider the following binary matrix*

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

The three rows of P are three 8-bit packets of data to be converted and encoded as follows:

1. *Convert to* 3×4 *matrix R where each row of R is the conversion of the 8-bit row of P to 4-symbol* 5*-ary with* $\mathbb{F}_5 = \{0,1,2,3,4\}$*. For example, considering base 2 and base 5 representations, we have* $(11111110)_2 = 254 = (2004)_5$*, so the first row of R is* $\begin{bmatrix} 2 & 0 & 0 & 4 \end{bmatrix}$.
2. *Encode to five 4-symbol* 5*-ary packets using* $\mathbb{F}_5 = \{0,1,2,3,4\}$ *and multiplication $C = AR$ with*

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

[4] We can use either A or $\tilde{A}$ for $(7,4)$ erasure coding in field $\mathbb{F}_7$, but the matrix $\tilde{A}$ uses the first four encoded packets as the original four packets.

where matrix multiplication is in $\mathbb{F}_5$. Note that any three rows of A are linearly independent (using field $\mathbb{F}_5$).

3. *Convert encoded matrix C row-by-row to five 10-bit binary encoded packets (a 5×10 binary matrix D), for overall efficiency of 8/10 = 0.8 (any three of the five 10-bit encoded packets can reconstruct the original data matrix P).*

a) Perform the conversion in this case to produce R, C, D. Show how the original matrix P can indeed be recovered from only the last three rows of D using the inverse of the last three rows of A. Explain how the inequality $2^8 < 5^4 < 2^{10}$ relates to efficiency, and how a large binary file can be taken in 8-bit chunks and encoded in this way with efficiency 0.8.

b) Suppose a large file is taken in 9-bit chunks. Explain how $2^9 < 5^4 < 2^{10}$ creates efficiency 0.9.

c) Suppose we take 100-bit chunks. Explain how $2^{100} < 5^{44} < 2^{103}$ creates efficiency $100/103 = 0.971$.

d) What would be the efficiency if we use 102-bit chunks?

EXERCISE 5.28. *(Probability of being invertible) Using the method described in Exercise 5.23c, compute the probability that a randomly formed $k \times k$ matrix with entries in a finite field is invertible. In particular, the field $\mathbb{F}$ has size $|\mathbb{F}| = f$; the columns of matrix A are independently chosen; each column is equally likely to be one of the f^k vectors in $\mathbb{F}^k$. The answer should be in terms of f and k. With k fixed, show that the probability converges to 1 as $f \to \infty$.*

EXERCISE 5.29. *(Null(A^k)) Let A be a square matrix. Suppose that Null(A^3) = Null(A^2).*

a) Prove that Null(A^4) = Null(A^2).

b) Prove that Null(A^k) = Null(A^2) for all integers $k \geq 2$.

EXERCISE 5.30. *(Rank) Suppose*

$$A = \begin{bmatrix} 1 & 0 & 4 \\ 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \end{bmatrix} \begin{bmatrix} d_1 & d_2 & d_3 & d_4 \\ d_5 & d_6 & d_7 & d_8 \\ 1 & 1 & 1 & 0 \end{bmatrix}$$

a) Explain why rank(A) $\in \{0, 1, 2\}$.

b) Give an example when rank(A) = 0.

c) Give an example when rank(A) = 1.

d) Give an example when rank(A) = 2.

EXERCISE 5.31. *Using $\mathbb{F} = \mathbb{R}$, find a specific square matrix A that satisfies $\dim Null(A^2) = 2 + \dim Null(A)$ (this relates to Section 5.10).*

EXERCISE 5.32. *(Reducing to a basis) Consider the following vectors in* $\mathbb{R}^3$*:*

$$\begin{bmatrix} 4 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 8 \\ 5 \\ 2 \end{bmatrix}, \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ -5 \\ 6 \end{bmatrix},$$

Reduce this to a linearly independent collection of vectors without changing their span. That is, throw away some vectors so that the remaining vectors are linearly independent and have the same span as before. What is the dimension of the span of these vectors?

EXERCISE 5.33. *(Reducing to a basis) Consider the following vectors in* $\mathbb{R}^3$*:*

$$\begin{bmatrix} -1 \\ 0 \\ 7 \end{bmatrix}, \begin{bmatrix} 8 \\ 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 1 \\ 8 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}$$

Reduce this to a linearly independent collection of vectors without changing their span. That is, throw away some vectors so that the remaining vectors are linearly independent and have the same span as before. What is the dimension of the span of these vectors?

EXERCISE 5.34. *(Reducing to a basis) Consider the following vectors in* $\mathbb{R}^3$*:*

$$\begin{bmatrix} -1 \\ 3 \\ 7 \end{bmatrix}, \begin{bmatrix} 3 \\ -9 \\ -21 \end{bmatrix}, \begin{bmatrix} 9 \\ -1 \\ 2 \end{bmatrix}, \begin{bmatrix} 29 \\ -9 \\ 8 \end{bmatrix}, \begin{bmatrix} 17 \\ 1 \\ 11 \end{bmatrix}, \begin{bmatrix} -23 \\ 17 \\ 31 \end{bmatrix}, \begin{bmatrix} 37 \\ -7 \\ 1 \end{bmatrix}$$

Reduce this to a linearly independent collection of vectors without changing their span. That is, throw away some vectors so that the remaining vectors are linearly independent and have the same span as before. What is the dimension of the span of these vectors?

EXERCISE 5.35. *(Growing a basis in* $\mathbb{R}^3$*) Define subspace* $S \subseteq \mathbb{R}^3$ *by*

$$S = \{(x, y, z) \in \mathbb{R}^3 : 2x + y - 3z = 0\}$$

a) What is $\dim(S)$*?*

b) Note that $(1, 1, 1) \in S$*. Extend* $(1, 1, 1)$ *to a basis for* S *by adding additional vectors.*

EXERCISE 5.36. *(Growing a basis in* $\mathbb{R}^4$*) Define subspace* $S \subseteq \mathbb{R}^4$ *by*

$$S = \{(x, y, z, w) \in \mathbb{R}^4 : x + y + z + w = 0\}$$

a) What is $\dim(S)$*?*

b) Note that $(1,1,1,-3) \in S$. *Extend* $(1,1,1,-3)$ *to a basis for S by adding additional vectors.*

EXERCISE 5.37. *(Growing a basis in* $\mathbb{R}^4$*) Define subspace* $S \subseteq \mathbb{R}^4$ *by*

$$S = \{(x,y,z,w) \in \mathbb{R}^4 : x + 2y + 3z + 4w = 0\}$$

a) What is $\dim(S)$*?*

b) Note that $(3,1,1,-2)$ *and* $(-7,0,1,1)$ *are linearly independent vectors in S. Extend* $\{(3,1,1,-2),(-7,0,1,1)\}$ *to a basis for S by adding additional vectors.*

EXERCISE 5.38. *(Growing a basis in* $\mathbb{R}^4$*) Define subspace* $S \subseteq \mathbb{R}^4$ *by*

$$S = \{(x,y,z,w) \in \mathbb{R}^4 : x + y + 2z + 2w = 0,\ x + z - w = 0\}$$

a) What is $\dim(S)$*?*

b) Note that $(2,2,-2,0) \in S$. *Extend* $(2,2,-2,0)$ *to a basis for S by adding additional vectors.*

EXERCISE 5.39. *(Growing a basis in an abstract way) Let S be a subspace of* $\mathbb{R}^{10}$*. Suppose we have vectors* $\{x,y,z\}$ *that are a basis for S (so* $\dim(S) = 3$*). Let v be a nonzero vector in S. How can we extend v to a basis for S?*

EXERCISE 5.40. *(Growing a basis in an abstract way) Let S be a subspace of* $\mathbb{F}^{42}$*, where* $\mathbb{F}$ *is some field. Suppose we have vectors* $\{x,y,z\}$ *that are a basis for S (so* $\dim(S) = 3$*). Let* v,w *be two linearly independent vectors in S. How can we extend* $\{v,w\}$ *to a basis for S?*

EXERCISE 5.41. *(Bounding the dimension) Suppose S is a subspace of* $\mathbb{F}^n$*. Suppose there is a positive integer k and vectors* $v_1,\ldots,v_k \in \mathbb{F}^n$ *such that*

$$S \subseteq Span\{v_1,\ldots,v_k\}$$

a) Show that $\dim(S) \leq k$.

b) Give an example of this situation where the inequality holds with strict inequality.

EXERCISE 5.42. *(Alternative proof of* $rank(AB) \leq rank(B)$*) Let A and B be matrices with entries in some field* $\mathbb{F}$*. Suppose the sizes are such that AB makes sense. Let* $r = rank(B)$.

a) Suppose $r = 0$*. Explain why B is an all-zero matrix, and* $rank(AB) \leq r$.

b) Suppose r is a positive integer. Let $\{w_1,\ldots,w_r\}$ *be a basis for* $Col(B)$*. Argue that*

$$Col(AB) \subseteq Span\{Aw_1,\ldots,Aw_r\}$$

Conclude that $rank(AB) \leq r$.

EXERCISE 5.43. *(Adding a vector already in the span of others) Let* $\{x_1,\ldots,x_k\}$ *be vectors in* $\mathbb{F}^m$*. Fix* $y \in Span\{x_1,\ldots,x_k\}$*. Show* $\{y,x_1,\ldots,x_k\}$ *are linearly dependent.*

EXERCISE 5.44. *(Removing a vector from a linearly dependent collection) Let $v_1, ..., v_k$ be vectors in $\mathbb{F}^m$. Define $S = Span\{v_1, ..., v_k\}$. Suppose $v_k \in Span\{v_1, ..., v_{k-1}\}$. We want to show that v_k can be removed without changing the span of what remains.*

a) Explain why $\{v_1, ..., v_k\}$ are linearly dependent.

b) Show that $Span\{v_1, ..., v_{k-1}\} = S$.

c) This problem assumed the last vector was a linear combination of the others. Explain why the result holds more generally: If any vector is a linear combination of the others, that vector can be removed without changing the span of the remaining vectors.

EXERCISE 5.45. *(A list of vectors where the first few are linearly independent) Let C be the following collection of vectors in $\mathbb{F}^m$:*

$$C = \{x_1, ..., x_k, y_1, ..., y_s\}$$

Define $S = Span\{x_1, ..., x_k, y_1, ..., y_s\}$. Suppose that

- *$\{x_1, ..., x_k\}$ are linearly independent.*
- *$\{x_1, ..., x_k, y_1, ..., y_s\}$ are linearly dependent.*

Prove that one of the y_i vectors is a linear combination of the other vectors in collection C. Hence, by Exercise 5.44, y_i can be removed from the collection C without changing the span of the remaining vectors.

EXERCISE 5.46. *(Wide matrix theorem via vector replacement) This exercise uses a* vector replacement *technique from Axler (2024) to prove the wide matrix theorem (Theorem 5.1). Suppose A is a $m \times n$ matrix with entries in some field $\mathbb{F}$. Suppose A is wide, so $m < n$. We want to prove $Null(A) \neq \{0\}$. Suppose the contrary holds, so $Null(A) = \{0\}$ (we reach a contradiction). Then the columns of A are linearly independent (recall Lemma 3.7). Let $x_1, ..., x_n$ be the (linearly independent) columns of A, so $x_i \in \mathbb{F}^m$ for $i \in \{1, ..., n\}$. Let I be the $m \times m$ identity matrix, having columns $e_1, ..., e_m \in \mathbb{F}^m$ that are the standard basis for $\mathbb{F}^m$. There are more vectors $x_1, ..., x_n$ than there are vectors $e_1, ..., e_m$.*

a) Show that every vector x_i is a linear combination of $\{e_1, ..., e_m\}$. [Just use the fact $x_i \in \mathbb{F}^m$.]

b) Step 1: Start with the collection of m vectors that span $\mathbb{F}^m$:

$$C_1 = \{e_1, e_2, ..., e_m\}$$

Add vector x_1 to the front of this collection to form a collection of $m + 1$ vectors that also span $\mathbb{F}^m$:

$$\tilde{C}_1 = \{x_1, e_1, e_2, ..., e_m\}$$

Use the results of Exercises 5.43-5.45 to show one of the e_i vectors can be removed from $\tilde{C}_1$, while the span of the remaining vectors is still $\mathbb{F}^m$.

c) Step $i \in \{2, \ldots, m\}$: At step i we start with the collection C_i of m vectors that span $\mathbb{F}^m$:

$$C_i = \{x_1, \ldots, x_{i-1}, e_{r[i,1]}, e_{r[i,2]}, \ldots, e_{r[i,m-(i-1)]}\}$$

where $r[i,1], \ldots, r[i, m-(i-1)]$ indexes the e_j vectors that have not been thrown away by the start of step i. Now add x_i to the front to form a collection of $m+1$ vectors that span $\mathbb{F}^m$:

$$\tilde{C}_i = \{x_1, \ldots, x_{i-1}, x_i, e_{r[i,1]}, e_{r[i,2]}, \ldots, e_{r[i,m-(i-1)]}\}$$

Use the results of Exercises 5.43-5.45 to show one of the $e_{r[i,j]}$ vectors can be removed from $\tilde{C}_i$, while the span of the remaining vectors is still $\mathbb{F}^m$.

d) At the end of step m we have thrown out all e_j vectors and we have the following collection of vectors that span $\mathbb{F}^m$:

$$\{x_1, \ldots, x_m\}$$

Since $m < n$, there is another vector $x_{m+1} \in \mathbb{F}^m$ that is not included in $\{x_1, \ldots, x_m\}$. Show this contradicts linear independence of $\{x_1, \ldots, x_n\}$.

EXERCISE 5.47. *(Counting the number of invertible 2×2 matrices) Let $\mathbb{F}$ be a finite field with f scalars. How many invertible 2×2 matrices are there with entries in $\mathbb{F}$?*

EXERCISE 5.48. *(Counting the number of invertible 3×3 matrices) Let $\mathbb{F}$ be a finite field with f scalars (where f is an integer and $f \geq 2$). How many invertible 3×3 matrices are there with entries in $\mathbb{F}$?*

EXERCISE 5.49. *(Counting the number of invertible $n \times n$ matrices) Let $\mathbb{F}$ be a finite field with f scalars (where f is an integer and $f \geq 2$). Fix $n \in \mathbb{N}$. How many invertible $n \times n$ matrices are there with entries in $\mathbb{F}$?*

EXERCISE 5.50. *Suppose A and B are 2×2 matrices with entries in $\mathbb{F}_2 = \{0, 1\}$ and*

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

where the matrix multiplication uses mod 2 arithmetic. How many different (A, B) combinations are there? Find all matrices A, B that work.

EXERCISE 5.51. *(Adding a vector to a linearly independent set) Let $\{x_1, \ldots, x_k\}$ be linearly independent vectors in some subspace S. Fix $y \in S$ such that $y \notin Span\{x_1, \ldots, x_k\}$. We want to show $\{x_1, \ldots, x_k, y\}$ are linearly independent. Suppose $\sum_{i=1}^{k} a_i x_i + a_{k+1} y = 0$ for some scalars a_i. Show $a_i = 0$ for all i by considering the cases $a_{k+1} = 0$ and $a_{k+1} \neq 0$.*

EXERCISE 5.52. *(Using dimension to establish a basis) Read the two paragraphs of Section 5.6. Then*

a) Prove Lemma 5.3(a).

b) Prove Lemma 5.3(b).

EXERCISE 5.53. *Let S_1 and S_2 be two subspaces of a vector space V. Suppose $S_1 \subseteq S_2$, and $\dim(S_1) = \dim(S_2) = k$ for some $k \in \mathbb{N}$. Show that $S_1 = S_2$.*

CHAPTER 6.

General vector spaces

This chapter presents the theory of general vector spaces. Most proofs are similar to the ones in previous chapters.

6.1. Axioms for a vector space over a field

Suppose V is a nonempty set of objects that shall be called *vectors*. Let $\mathbb{F}$ be a scalar field. Suppose there are operations of *vector addition* and *scalar multiplication,* so that adding two vectors in V yields another vector in V, and multiplying a vector in V by a scalar yields another vector in V:

$$x \in V, y \in V \implies x + y \in V$$
$$a \in \mathbb{F}, x \in V \implies ax \in V$$

The set V, together with the field $\mathbb{F}$ and the operations of vector addition and scalar multiplication, is called a *vector space over field* $\mathbb{F}$ if the following eight axioms hold:

1. Vector addition is commutative: $x + y = y + x$ for all $x, y \in V$.
2. Vector addition is associative: $(x + y) + z = x + (y + z)$ for all $x, y, z \in V$.
3. Existence of vector 0: There is a vector $0 \in V$ (called the zero vector) such that $x + 0 = x$ for all $x \in V$.
4. Existence of additive inverse vector: For every $x \in V$ there is a $-x \in V$ such that $x + -x = 0$.

5. $1 \cdot x = x$ for all $x \in V$ (where 1 is the multiplicative identity scalar in $\mathbb{F}$).
6. $(ab)x = a(bx)$ for all $a, b \in \mathbb{F}$ and $x \in V$.
7. $a(x + y) = ax + ay$ for all $a \in \mathbb{F}$ and $x, y \in V$.
8. $(a + b)x = ax + bx$ for all $a, b \in \mathbb{F}$ and $x \in V$.

For $x, y \in V$, we define *subtraction* of vector y as adding $-y$:

$$x - y = x + -y$$

The zero vector $0 \in V$ is not the same as the scalar $0 \in \mathbb{F}$. The above axioms can be used to show that 0 is the *unique* vector that can be added to another vector without changing that other vector. Indeed, fix $x \in V$. Suppose there is some $z \in V$ such that $x + z = x$. We want to show z must be the vector 0. We have

$$\begin{aligned}
0 &= x + -x && \text{[existence of } -x \text{ (Axiom 4)]} \\
&= (x + z) + -x && \text{[substitute } x = x + z\text{]} \\
&= (z + x) + -x && \text{[commutativity (Axiom 1)]} \\
&= z + (x + -x) && \text{[associativity (Axiom 2)]} \\
&= z + 0 && \text{[Axiom 4]} \\
&= z && \text{[Axiom 3]}
\end{aligned}$$

LEMMA 6.1. *(Basic consequences of the axioms) Let V be a vector space over a field $\mathbb{F}$. Then*

- $0x = 0$ *for all* $x \in V$. *(Multiplying any vector by the scalar* 0 *yields the vector* 0.*)*
- $a0 = 0$ *for all* $a \in \mathbb{F}$. *(Multiplying the vector* 0 *by any scalar yields the vector* 0.*)*
- $(-1)x = -x$ *for all* $x \in V$. *(Multiplying any vector x by -1 yields $-x$.)*
- *If* $a \in \mathbb{F}$ *and* $x \in V$ *then:* $(ax = 0) \iff$ (*either* $a = 0$ *or* $x = 0$)
- *For all* $x, y \in V$ *we have* $x = x + y - y$. *(The famous add-subtract technique.)*

PROOF. The proof is similar to proofs in Chapter 1.5. Details are left to Exercise 6.32. □

6.1.1. Terminology

- A vector space over the field $\mathbb{R}$ is called a *real vector space*.
- A vector space over the field $\mathbb{C}$ is called a *complex vector space*.
- $V = \{0\}$ is called the *trivial vector space*.
- A vector space is *nontrivial* if it has at least two distinct vectors.

6.1.2. Finite versus infinite sums

Suppose V is a vector space over a field $\mathbb{F}$. Let $\{v_1, v_2, v_3, \ldots\}$ be an infinite sequence of vectors in V. The axioms for a vector space imply that the sum of any two vectors in V is again a vector in V, so

$$v_1 + v_2 \in V$$

Therefore

$$v_1 + v_2 + v_3 = \underbrace{(v_1 + v_2)}_{\in V} + v_3 \in V$$

$$v_1 + v_2 + v_3 + v_4 = \underbrace{(v_1 + v_2 + v_3)}_{\in V} + v_4 \in V$$

By induction it follows that for all positive integers n we have

$$\sum_{i=1}^{n} v_i \in V$$

Conclusion: The axioms for a vector space ensure that the sum of any finite number of vectors in V is again a vector in V. Similarly, if $\{a_1, a_2, a_3, \ldots\}$ is a sequence of scalars, then $a_i v_i \in V$ for all $i \in \mathbb{N}$, so $\sum_{i=1}^{n} a_i v_i \in V$ for all $n \in \mathbb{N}$.

What about infinite sums?

$$\sum_{i=1}^{\infty} v_i = ?? \quad , \quad \sum_{i=1}^{\infty} a_i v_i = ??$$

The axioms for a vector space cannot and do not define such infinite sums. Linear algebra deals with *finite sums*. See also Exercises 1.24, 6.29, and the definition of *Hamel basis* in Exercise 6.31.

[Of course, using the real field $\mathbb{R}$ or the complex field $\mathbb{C}$, it is often possible to define the *distance* between vectors. Then one can talk about *limits*. With this added machinery, infinite sums can be defined for certain vector spaces and when certain convergence criteria hold. Nevertheless, linear algebra primarily deals with finite sums. The one place an infinite sum is used in this text, aside from the Exercises 1.24 and 6.29, is in the definition of the *matrix exponential* function e^{At} in Chapter 14.]

6.2. Linear combinations

Let V be a vector space over a field $\mathbb{F}$.

DEFINITION 6.1. *Let D be a nonempty subset of V. A vector x is a* linear combination *of*

vectors in D if

$$x = \sum_{i=1}^{k} a_i x_i$$

for some positive integer k, some vectors $x_1, \ldots, x_k \in D$*, and some scalars* $a_1, \ldots, a_k \in \mathbb{F}$.

The scalars used to form a linear combination are often called *weights*, so a linear combination of vectors in D is a weighted sum of a finite number of vectors from D. The set D in the above definition is just some nonempty subset of V. It is not required to have any special structure. For example, D might contain an infinite number of vectors. However, by definition, a linear combination is a *finite* sum over k terms, for some $k \in \mathbb{N}$.

Suppose $\{v_1, v_2, v_3, \ldots\}$ is a countably infinite list of vectors in V. If $\{a_1, a_2, a_3, \ldots\}$ is a countably infinite list of scalars, the summation $\sum_{i=1}^{\infty} a_i v_i$ is not necessarily a linear combination of $\{v_1, v_2, v_3, \ldots\}$ because: (i) Unless additional structure is given, infinite sums are undefined; (ii) Even if we have structure for which $\sum_{i=1}^{\infty} a_i v_i$ is defined as some other vector $w \in V$, it is only a linear combination of $\{v_1, v_2, v_3, \ldots\}$ if it can be written as $w = \sum_{i=1}^{k} \tilde{a}_i v_i$ for some $k \in \mathbb{N}$ and some scalars $\tilde{a}_i$ for $i \in \{1, \ldots, k\}$. See a related example in Exercise 6.29.

DEFINITION 6.2. *Given a collection of vectors* $\{x_1, \ldots, x_k\}$ *in* V*, we define* $Span\{x_1, \ldots, x_k\}$ *as the set of all linear combinations of* $\{x_1, \ldots, x_k\}$.

6.3. Subspaces

Let V be a vector space over a field $\mathbb{F}$.

DEFINITION 6.3. *A set* $S \subseteq V$ *is a* subspace of V *if the following three properties hold:*

(i) $0 \in S$*;*

(ii) $x + y \in S$ *whenever* $x, y \in S$*;*

(iii) $ax \in S$ *whenever* $x \in S$ *and* $a \in \mathbb{F}$.

Consequences of this definition:

- An equivalent definition: Let S be a subset of V. Then S is a subspace of V if and only if S is nonempty and $ax + by \in S$ whenever $x, y \in S$ and $a, b \in \mathbb{F}$. [The proof is similar to the proof of Lemma 3.1.]
- V itself is a subspace of V.
- The single-element set $\{0\}$ is a subspace of V. It is called the *trivial subspace*.
- If S is a subspace of V, then any linear combination of vectors in S is itself a vector in S. [This can formally be shown by induction, similar to the proof of Lemma 3.2.]
- If $\{x_1, \ldots, x_k\}$ is a collection of vectors in V, the set $Span\{x_1, \ldots, x_k\}$ is a subspace of V.

[The proof is similar to the one given for vector space $\mathbb{F}^n$ in Chapter 3.2.2.]

- If D is a nonempty subset of V, then the set of all linear combinations of vectors in D is a subspace of V. [The proof is similar to the proof of the previous bullet.]
- The intersection of an arbitrary (possibly uncountably infinite) number of subspaces of V is again a subspace of V. [The proof is not significantly different from the proof for the intersection of two subspaces given in Chapter 3.2.2.]

Let V be a vector space over a field $\mathbb{F}$. It can be shown that every subspace $S \subseteq V$ is itself a legitimate vector space over field $\mathbb{F}$. This is useful: If the axioms are known to hold for V, then they necessarily hold for any subspace of V. New vector spaces can be formed from old ones just by using subspaces.

6.4. Example vector spaces

1. (n-tuples) Fix $\mathbb{F}$ as any scalar field and fix n as a positive integer. Define

$$V = \mathbb{F}^n = \{(x_1, ..., x_n) : x_i \in \mathbb{F} \quad \forall i \in \{1, ..., n\}\}$$

Using the standard definition of addition and scalar multiplication in $\mathbb{F}^n$, it is easy to show that $\mathbb{F}^n$ is a vector space over the field $\mathbb{F}$. This is done by showing that all axioms hold. For example, Axiom 1 holds because we already know the sum of two vectors in $\mathbb{F}^n$ is commutative:

$$\begin{aligned}(x_1, ..., x_n) + (y_1, ..., y_n) &= (x_1 + y_1, ..., x_n + y_n) \\ &= (y_1 + x_1, ..., y_n + x_n) \\ &= (y_1, ..., y_n) + (x_1, ..., x_n)\end{aligned}$$

Axiom 2 (associativity of vector sums) can be similarly shown. For Axiom 3, we define the zero vector as the all-zero n-tuple

$$0 = (0, 0, ..., 0)$$

Then it is clear that for all $x = (x_1, ..., x_n) \in \mathbb{F}^n$ we have

$$\begin{aligned}x + 0 &= (x_1, ..., x_n) + (0, ..., 0) \\ &= (x_1 + 0, ..., x_n + 0) \\ &= (x_1, ..., x_n) = x\end{aligned}$$

For Axiom 4, given $x = (x_1, \ldots, x_n)$ we define $-x = (-x_1, \ldots, -x_n)$, so $x + -x = 0$. The remaining Axioms 5, 6, 7, 8 can be similarly shown.

[Note: We do not need to assign a vertical or horizontal orientation to vectors in $\mathbb{F}^n$. However, when used in matrix multiplication, we follow the convention that a vector $x \in \mathbb{F}^n$ is viewed as a column vector (a $n \times 1$ matrix), so the transpose $x^\top$ makes sense and is a row vector (a $1 \times n$ matrix).]

2. ($m \times n$ matrices) Fix $\mathbb{F}$ as a scalar field. Fix $m, n \in \mathbb{N}$. Let V be the set of all $m \times n$ matrices with entries in $\mathbb{F}$, often denoted $\mathbb{F}^{m \times n}$:

$$V = \mathbb{F}^{m \times n} = \{m \times n \text{ matrices with entries in } \mathbb{F}\}$$

It is not surprising that $\mathbb{F}^{m \times n}$ is a vector space over the field $\mathbb{F}$. For example, we already know that matrix summation is commutative and associative, so Axioms 1 and 2 hold. The vector 0 is defined as the all-zero matrix of size $m \times n$, from which it is clear that Axiom 3 holds. It is not difficult to verify that the remaining axioms (Axioms 4-8) also hold. The vector space $\mathbb{F}^{m \times n}$ is similar to the vector space $\mathbb{F}^{mn}$, which takes the $m \times n$ matrix and splices all columns together in a long column vector of size mn. However, the rectangular structure of matrices in $\mathbb{F}^{m \times n}$ is useful when we want to use features of matrix transposes and matrix multiplication. For example from the vector space $\mathbb{F}^{m \times n}$, we can define a subspace S as follows: Fix a particular square matrix $B \in \mathbb{F}^{n \times n}$ and define

$$S = \{A \in \mathbb{F}^{m \times n} : AB = 0\}$$

where 0 denotes the all-zero matrix of size $m \times n$. Exercise 6.35 shows that S is indeed a subspace of $\mathbb{F}^{m \times n}$.

3. (Infinite sequences) Fix $\mathbb{F}$ as a scalar field. Define

$$V = \mathbb{F}^{\mathbb{N}} = \{(x_1, x_2, x_3, \ldots) : x_i \in \mathbb{F} \quad \forall i \in \mathbb{N}\}$$

This is the set of all infinite sequences of scalars in $\mathbb{F}$. So

$$(x \in \mathbb{F}^{\mathbb{N}}) \iff (x = (x_1, x_2, x_3, \ldots) \quad \text{for some scalars } x_i \in \mathbb{F})$$

As convenient shorthand, we can denote a particular sequence $x \in \mathbb{F}^{\mathbb{N}}$ by (x_i). Define vector addition and scalar multiplication by

$$(x_i) + (y_i) = (x_i + y_i) \quad \forall x, y \in \mathbb{F}^{\mathbb{N}}$$
$$\theta(x_i) = (\theta x_i) \quad \forall x \in \mathbb{F}^{\mathbb{N}}, \theta \in \mathbb{F}$$

Define the vector 0 as the all-zero sequence $0 = (0, 0, 0, \ldots)$. It can be shown that $\mathbb{F}^{\mathbb{N}}$ is a

vector space over field $\mathbb{F}$. For example, to show Axiom 1, fix $(x_i), (y_i) \in \mathbb{F}^{\mathbb{N}}$ and observe

$$(x_i) + (y_i) = (x_i + y_i) = (y_i + x_i) = (y_i) + (x_i)$$

The remaining axioms (Axioms 2-8) can be similarly shown. Regarding Axiom 4, for each $(x_i) \in \mathbb{F}^{\mathbb{N}}$ define $-(x_i)$ as the vector $(-x_i)$, for which it is clear that $(x_i) + -(x_i) = 0$.

4. (Real-valued functions) Fix $\mathbb{F} = \mathbb{R}$. Define V as the set of all functions $f : \mathbb{R} \to \mathbb{R}$. For example, the following functions f_1, f_2, f_3, f_4 are in set V:

$$f_1(t) = t + t^2 \quad \forall t \in \mathbb{R} \tag{6.1}$$

$$f_2(t) = \cos(t) - e^t \quad \forall t \in \mathbb{R} \tag{6.2}$$

$$f_3(t) = \begin{cases} 2 & \text{if } t < 1.5 \\ -3 & \text{if } t \geq 1.5 \end{cases} \tag{6.3}$$

$$f_4(t) = \begin{cases} 0 & \text{if } t < 0 \\ e^{-2t} & \text{if } t \geq 0 \end{cases} \tag{6.4}$$

The functions f_1, f_2 are continuous. The functions f_3, f_4 are not continuous. Two functions $f, g \in V$ are said to be *equal* if their function values are the same for all $t \in \mathbb{R}$:

$$(f = g) \iff (f(t) = g(t) \quad \forall t \in \mathbb{R}) \tag{6.5}$$

For example, with f_1 and f_2 defined as above, we have $f_1(0) = f_2(0)$ (so the function values happen to be equal at time $t = 0$) but $f_1 \neq f_2$ (because they are different functions). The fact that $f_1 \neq f_2$ can be established by finding *any* time $t \in \mathbb{R}$ for which $f_1(t) \neq f_2(t)$. We can also see that $f_1 \neq f_2$ by observing

$$\lim_{t \to \infty} f_1(t) = \infty$$
$$\lim_{t \to \infty} f_2(t) = -\infty$$

To allow V to be a real vector space, define addition and scalar multiplication according to standard sums and scalings of real-valued functions. Formally, if $f, g \in V$ and $\theta \in \mathbb{R}$ we define $f + g$ and θf as functions in V with values at each time $t \in \mathbb{R}$ given by

$$(f + g)(t) = f(t) + g(t) \quad \forall t \in \mathbb{R} \tag{6.6}$$

$$(\theta f)(t) = \theta f(t) \quad \forall t \in \mathbb{R} \tag{6.7}$$

where the notation $(f + g)(t)$ and $(\theta f)(t)$ evaluates the functions $f + g$ and θf at the

particular time t. Then V is a real vector space. This can be proven by verifying that a linear combination of two functions in V is again a function in V, then showing all eight axioms hold. To establish Axioms 1 and 2, fix $f, g, h \in V$ and observe

$$f + g = g + f \quad \text{(Axiom 1)}$$
$$(f + g) + h = f + (g + h) \quad \text{(Axiom 2)}$$

Note that functions f and g represent functions over all time $t \in \mathbb{R}$, so $f + g$ is also a function over all $t \in \mathbb{R}$. Thus, Axiom 1 means

$$f(t) + g(t) = g(t) + f(t) \quad \forall t \in \mathbb{R}$$

For Axiom 3, define the zero vector 0 as the always-zero function $z(t) = 0$ for all $t \in \mathbb{R}$. Then it is clear that $f + 0 = f$ (Axiom 3). For Axiom 4, for each function $f \in V$ we define $-f$ as the function $g : \mathbb{R} \to \mathbb{R}$ given by $g(t) = -f(t)$ for all $t \in \mathbb{R}$. Then it is clear that $f + -f = 0$. It is easy to see that Axioms 5-8 also hold (for example, $1 \cdot f = f$ for all functions $f \in V$). So V is a real vector space.

Now define $S_1 \subseteq V$ and $S_2 \subseteq V$ by

$$S_1 = \{f \in V : f \text{ is continuous}\}$$
$$S_2 = \left\{f \in V : \int_{-\infty}^{\infty} f(t)^2 dt < \infty\right\}$$

With f_1, f_2, f_3, f_4 defined in (6.1)-(6.4), we have

- $f_1, f_2 \in S_1$ [since functions f_1 and f_2 are continuous]
- $f_3, f_4 \notin S_1$ [since functions f_3 and f_4 are not continuous]
- $f_1, f_2, f_3 \notin S_2$ [since, for example, $\int_{-\infty}^{\infty} f_3(t)^2 dt = \int_{-\infty}^{1.5} 4dt + \int_{1.5}^{\infty} 9dt = \infty + \infty = \infty$]
- $f_4 \in S_2$ [since $\int_{-\infty}^{\infty} f_4(t)^2 dt = \int_0^{\infty} e^{-4t} dt = 1/4 < \infty$]

Since the zero function is continuous, and the linear combination of two continuous functions is again a continuous function, the set S_1 is a subspace of V. Thus, S_1 is itself a valid vector space. It can also be shown that S_2 is a subspace of V (see Exercise 6.3). The set S_2 is the real vector space of *square integrable functions* $f : \mathbb{R} \to \mathbb{R}$.

5. (Polynomial functions) Fix $\mathbb{F} = \mathbb{R}$. As in the previous bullet, define V as the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$. Define V_3 as the subset of V consisting of polynomial functions of degree less than or equal to 3:

$$V_3 = \{f \in V : f(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \text{ for some } a_0, a_1, a_2, a_3 \in \mathbb{R}\}$$

It can be shown that V_3 is a subspace of V. Thus, V_3 is itself a real vector space. Now define $B \subseteq V_3$ as the set of all polynomial functions of degree *exactly* 3. While B is a subset of V_3, observe that B is *not* a subspace of V_3 because $t^3 \in B$, but $0t^3 = 0 \notin B$.

6. (Functions on any domain) Let $\mathbb{F} = \mathbb{R}$. Let D be any nonempty set. Fix $k \in \mathbb{N}$. Define V as the set of all functions $f : D \to \mathbb{R}^k$. Each function $f \in D$ maps an element $d \in D$ to a k-tuple $f(d) \in \mathbb{R}^k$. Such functions in V can be added and multiplied by scalars to produce other functions in V. It can be shown that V is a real vector space. The zero vector is the always-zero function $f(d) = 0 \in \mathbb{R}^k$ for all $d \in D$.

   For example, let $D = \{\text{red, green, blue}\}$. Let $k = 1$. Define $f : D \to \mathbb{R}$ and $g : D \to \mathbb{R}$ by

$$\begin{aligned} f(\text{red}) = 1 \quad , \quad f(\text{green}) = 2.2 \quad , \quad f(\text{blue}) = 0 \\ g(\text{red}) = 3 \quad , \quad f(\text{green}) = 3 \quad , \quad f(\text{blue}) = 1 \end{aligned}$$

   Define $h = f + g$ and $r = 5.1f - g$. Then

$$\begin{aligned} h(\text{red}) = 4 \quad , \quad h(\text{green}) = 5.2 \quad , \quad h(\text{blue}) = 1 \\ r(\text{red}) = 2.1 \quad , \quad r(\text{green}) = 8.22 \quad , \quad r(\text{blue}) = -1 \end{aligned}$$

   For these examples, it is clear that $f + g = g + f$ (required by Axiom 1); $(f + g) + h = f + (g + h)$ (required by Axiom 2); $f + 0 = f$ (required by Axiom 3); and so on.

7. (Random variables) Let $(\Omega, \mathcal{F}, P)$ be a probability space: Ω is a nonempty set (called the sample space); $\mathcal{F}$ is a sigma algebra on Ω (called the set of events); $P : \mathcal{F} \to [0, 1]$ is a probability measure. Define $\mathbb{F} = \mathbb{R}$ and define V as the set of all random variables $X : \Omega \to \mathbb{R}$. It can be shown that V is a real vector space. This is done by verifying existence of vector addition and scalar multiplication (the sum of two random variables is again a random variable, multiplying a random variable by a real number is again a random variable) and then showing the 8 axioms hold. The zero vector is the always-zero random variable $Z(\omega) = 0$ for all $\omega \in \Omega$. A useful subspace of V is the set of all random variables that have finite second moments (see Exercise 6.36).

6.5. Finite and infinite dimensional subspaces

Let V be a vector space over a field $\mathbb{F}$.

DEFINITION 6.4. *A subspace $S \subseteq V$ is* finite dimensional*, written* $\dim(S) < \infty$*, if S is spanned by a finite number of vectors. That is, $S = Span\{x_1, ..., x_k\}$ for some positive integer k and some vectors $x_1, ..., x_k \in S$.*

DEFINITION 6.5. *A subspace $S \subseteq V$ is* infinite dimensional*, written* $\dim(S) = \infty$*, if it is not*

finite dimensional.

For example, let V be the real vector space of all functions $f : \mathbb{R} \to \mathbb{R}$, and let $S \subseteq V$ be the subspace of polynomial functions. The set S is indeed a subspace because a linear combination of two polynomials is again a polynomial. We now prove that $\dim(S) = \infty$: Suppose S is the span of a finite number of polynomials, so $S = Span\{p_1(t), p_2(t), ..., p_m(t)\}$ for some positive integer m and some polynomials $p_1(t), ..., p_m(t)$ (we reach a contradiction). Let d be the maximum degree of any of these m polynomials. Then $t^{d+1} \in S$ but $t^{d+1} \notin Span\{p_1(t), p_2(t), ..., p_m(t)\}$ because the linear combination of polynomials of degree at most d is again a polynomial of degree at most d. This contradicts the assumption $S = Span\{p_1(t), p_2(t), ..., p_m(t)\}$. Thus, $\dim(S) = \infty$.

Remarkably, it is more difficult to prove that V itself is infinite dimensional.[1] We know $S \subseteq V$, but we do not yet have a general way to prove that an infinite dimensional subspace cannot be contained by a finite dimensional subspace (this is done in the last paragraph of Section 6.6 using the concept of *linear independence*). Exercise 6.30 shows that for any field $\mathbb{F}$, the vector space $\mathbb{F}^{\mathbb{N}}$ of all infinite sequences of scalars in $\mathbb{F}$ is not only infinite dimensional (so it cannot be spanned by any finite list of vectors in $\mathbb{F}^{\mathbb{N}}$), but it is so large that it cannot be spanned by any countably infinite list of vectors in $\mathbb{F}^{\mathbb{N}}$.

6.6. Linear independence, basis, and dimension

Let V be a vector space over a field $\mathbb{F}$. Let k be a positive integer. A collection of vectors $\{x_1, ..., x_k\}$ in V is said to be *linearly independent* if

$$\left(\sum_{i=1}^{k} a_i x_i = 0\right) \implies (a_i = 0 \quad \forall i)$$

where $a_1, ..., a_k$ are assumed to be scalars in $\mathbb{F}$. In other words, the only way to linearly combine the vectors $\{x_1, ..., x_k\}$ to yield vector 0 is to use zero for all scalar weights. Vectors $\{x_1, ..., x_k\}$ are said to be *linearly dependent* if they are not linearly independent.

Consequences of this definition (assume k is a positive integer):

- If $\{x_1, ..., x_k\}$ are linearly independent, then any $y \in Span\{x_1, ..., x_k\}$ has the form $y = \sum_{i=1}^{k} a_i x_i$ for some *unique* scalar weights $(a_1, ..., a_k)$.
- $\{x_1, ..., x_k\}$ are linearly dependent $\iff$ either $x_1 = 0$, or there is an $i \in \{2, ..., k\}$ such that $x_i \in Span\{x_1, ..., x_{i-1}\}$.

[1] A direct proof that $\dim(V) = \infty$ starts by assuming $V = Span\{f_1, ..., f_m\}$ for some $m \in \mathbb{N}$ and some $f_1, ..., f_m \in V$, then reaches a contradiction by building a function $g : \mathbb{R} \to \mathbb{R}$ that is not in $Span\{f_1, ..., f_m\}$. Such g can be built by sampling the f_i functions at a suitably large number of points in their domain.

- $\{x_1, ..., x_k\}$ are linearly independent $\iff$ $x_1 \neq 0$ and for all $i \in \{2, \ldots, k\}$ we have $x_i \notin Span\{x_1, ..., x_{i-1}\}$.
- If v is a nonzero vector in V, then $\{v\}$ is linearly independent.
- If a list of vectors $\{x_1, ..., x_k\}$ contains the vector 0, or contains a redundantly repeated vector, or contains any vector that is a linear combination of the remaining vectors, then $\{x_1, ..., x_k\}$ is linearly dependent.

Let S be a subspace of V. A collection of vectors $\{y_1, ..., y_m\}$ is said to *span* S if

$$Span\{y_1, ..., y_m\} = S$$

If a collection of vectors span S, then all vectors in that collection must be in S.

THEOREM 6.1. *(independent-set-spanning-set comparison for a general vector space) Let V be a vector space over a field $\mathbb{F}$. Suppose S is a subspace of V and:*

(i) $\{x_1, ..., x_k\}$ are linearly independent vectors in S;

(ii) $\{y_1, ..., y_m\}$ are vectors that span S.

Then $k \leq m$.

PROOF. Suppose $k > m$ (we reach a contradiction). Fix $j \in \{1, \ldots, k\}$. Since $x_j \in S$ and $\{y_1, \ldots, y_m\}$ span S, we have

$$x_j = \sum_{i=1}^{m} a_{ij} y_i \tag{6.8}$$

for some scalars a_{ij}. Let $A = (a_{ij})$ be the $m \times k$ matrix composed of these scalars. Since $k > m$, A is a wide matrix so Theorem 5.1 ensures there is a nonzero vector $v \in \mathbb{F}^k$ such that $Av = 0$. That is, with $v = (v_1, ..., v_k) \neq 0$ we have

$$\sum_{j=1}^{k} a_{ij} v_j = 0 \quad \forall i \in \{1, ..., m\} \tag{6.9}$$

Now we use the v_j coefficients (not all of which are zero) to form a weighted sum of $\{x_1, ..., x_k\}$ that gives zero, which contradicts linear independence of $\{x_1, ..., x_k\}$. We have:

$$\sum_{j=1}^{k} v_j x_j \overset{(a)}{=} \sum_{j=1}^{k} v_j \left[\sum_{i=1}^{m} a_{ij} y_i \right] = \sum_{i=1}^{m} y_i \underbrace{\left[\sum_{j=1}^{k} a_{ij} v_j \right]}_{0} \overset{(b)}{=} 0$$

where (a) holds by (6.8) and (b) holds by (6.9). □

Let S be a subspace of V. Let k be a positive integer. A collection of vectors $\{x_1, ..., x_k\}$ in S is said to be a *basis for S* if

- $\{x_1, ..., x_k\}$ are linearly independent.
- $\{x_1, ..., x_k\}$ span S.

LEMMA 6.2. *If S is a nontrivial and finite dimensional subspace, then it has a basis.*

PROOF. Since S is finite dimensional, there is a finite collection of vectors $\{x_1, ..., x_m\}$ that span S. Not all vectors in this collection are zero because S is nontrivial. Without loss of generality, assume $\{x_1, ..., x_m\}$ are all nonzero vectors (else, just throw away the zeros, which does not change the span of what remains). Now we repeatedly throw away any single vector in our collection that is a linear combination of the others (which does not change the span of what remains). We stop when our collection is linearly independent. This procedure is guaranteed to terminate because, in the extreme case, we reduce the collection to just one nonzero vector, which by itself is linearly independent and hence must be a basis for S. □

The general independent-set-spanning-set-comparison theorem can be used to show that if $\{x_1, ..., x_k\}$ and $\{y_1, ..., y_m\}$ are any two bases of a subspace S, then $k = m$. [The reasoning is the same as that in Chapter 5.3.] When a subspace S is finite dimensional, the *dimension* of S, written $\dim(S)$, is defined as the number of vectors in a basis for S. The trivial subspace $S = \{0\}$ has no basis and is said to have dimension 0. Thus

$$(S \text{ is finite dimensional}) \iff (\dim(S) < \infty) \iff \dim(S) \in \{0, 1, 2, 3, ...\}$$

This text treats the concept of *basis for a subspace S* only when S is a finite dimensional subspace. Thus, for our purposes, a basis is always a finite collection of vectors. [Within mathematics, there are several nonequivalent ways to define a basis for an infinite dimensional subspace. One way is to use the *Hamel basis* definition, explored in Exercise 6.31.]

When a subspace S is infinite dimensional we say $\dim(S) = \infty$. The following procedure can be used to extract an arbitrarily large collection of linearly independent vectors from an infinite dimensional subspace S: Choose any nonzero vector in S and call it x_1; Choose any vector in S that is not in $Span\{x_1\}$ and call it x_2; Choose any vector in S that is not in $Span\{x_1, x_2\}$ and call it x_3. And so on. At each step k, this procedure ends with a collection of vectors $\{x_1, ..., x_k\}$ that is linearly independent in S. This procedure can only terminate at some step k if $S = Span\{x_1, ..., x_k\}$, which is impossible because $\dim(S) = \infty$. Therefore, the procedure never terminates and it produces an infinite list of vectors $\{x_1, x_2, x_3, ...\}$ in S such that for all $k \in \mathbb{N}$, the vectors $\{x_1, ..., x_k\}$ are linearly independent.

This fact can be used to prove that the real vector space V of all functions $f : \mathbb{R} \to \mathbb{R}$ is infinite dimensional (this question was posed at the end of Section 6.5): Suppose V is

spanned by a finite number m of functions in V. We know that the set S of all polynomial functions in V is an infinite dimensional subspace of V, so we must be able to extract $m+1$ linearly independent vectors from S, these are then $m+1$ linearly independent vectors of V, contradicting the independent-set-spanning-set comparison theorem. In general, this theorem can be used to prove the following: If S_1 and S_2 are two subspaces of V, then

$$S_1 \subseteq S_2 \implies \dim(S_1) \leq \dim(S_2)$$

6.7. Using dimension to establish a basis

Let V be a vector space over a field $\mathbb{F}$.

LEMMA 6.3. *Let S be a subspace of V. Suppose $\dim(S) = k$ for some $k \in \mathbb{N}$. Then*
a) If $\{x_1, \ldots, x_k\}$ are linearly independent vectors in S, then they are a basis for S.
b) If $\{x_1, \ldots, x_k\}$ span S, then they are a basis for S.

PROOF. See Exercise 6.42. □

LEMMA 6.4. *Suppose S_1, S_2 are subspaces of V such that $\dim(S_2) < \infty$ and $S_1 \subseteq S_2$. Then*

$$(S_1 = S_2) \iff (\dim(S_1) = \dim(S_2))$$

$$(S_1 \subsetneq S_2) \iff (\dim(S_1) < \dim(S_2))$$

where $S_1 \subsetneq S_2$ means that S_1 is a strict subset *of S_2, so all elements of S_1 are in S_2, but there is at least one element of S_2 that is not in S_1.*

PROOF. See Exercise 6.43. □

It is interesting to note that the conclusion of Lemma 6.4 can fail if the condition $\dim(S_2) < \infty$ is removed. In particular, it is possible to have subspaces S_1 and S_2 that satisfy $S_1 \subseteq S_2$ and $\dim(S_1) = \dim(S_2) = \infty$, but $S_1 \neq S_2$ (see Exercise 6.44).

6.8. Examples

6.8.1. Linear independence of $\{\cos(t), t, t^2\}$

Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$. To prove that three given functions $f_1, f_2, f_3 \in V$ are linearly independent, we start the proof:

Suppose there are scalars $a_1, a_2, a_3 \in \mathbb{R}$ such that $a_1 f_1 + a_2 f_2 + a_3 f_3 = 0$. We want to show $a_1 = a_2 = a_3 = 0$.

An equivalent way to start the proof is

Suppose there are scalars $a_1, a_2, a_3 \in \mathbb{R}$ such that

$$a_1 f_1(t) + a_2 f_2(t) + a_3 f_3(t) = 0 \quad \forall t \in \mathbb{R} \tag{6.10}$$

We want to show $a_1 = a_2 = a_3 = 0$.

This is equivalent because

$$(a_1 f_1 + a_2 f_2 + a_3 f_3 = 0) \iff (a_1 f_1(t) + a_2 f_2(t) + a_3 f_3(t) = 0 \quad \forall t \in \mathbb{R})$$

To complete the proof, we need to use equation (6.10) to conclude that $a_1 = a_2 = a_3 = 0$. This can often be done by sampling that equation at three particular times, say, times $t = 0, t = 1, t = 7.7$, to obtain three linear equations with three unknowns a_1, a_2, a_3:

$$\begin{aligned} a_1 f_1(0) + a_2 f_2(0) + a_3 f_3(0) &= 0 \\ a_1 f_1(1) + a_2 f_2(1) + a_3 f_3(1) &= 0 \\ a_1 f_1(7.7) + a_2 f_2(7.7) + a_3 f_3(7.7) &= 0 \end{aligned}$$

Equivalently

$$\underbrace{\begin{bmatrix} f_1(0) & f_2(0) & f_3(0) \\ f_1(1) & f_2(1) & f_3(1) \\ f_1(7.7) & f_2(7.7) & f_3(7.7) \end{bmatrix}}_{A} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

If we are *lucky* then the three times at which we have sampled equation (6.10) yield three linearly independent equations, so the 3×3 matrix A defined in the underbrace above is invertible, allowing us to conclude $(a_1, a_2, a_3) = (0, 0, 0)$. [Formally, we only define "linearly independent" for a list of vectors. However, it is common to call a set of consistent equations "linearly independent" if each equation offers new information, so no equation can be inferred as a consequence of the others. For example, the two equations $a + b = 1$, $2a + 2b = 2$ are not linearly independent since the second can be inferred from the first. Typically this terminology is used for k linear equations of the form $Ax = b$, where A is a $k \times n$ matrix, in which case the k equations are linearly independent if and only if the k rows of A are linearly independent, which holds if and only if Gaussian elimination produces a pivot on each row.]

If we are *unlucky* then the times we have sampled equation (6.10) do not lead to linearly independent equations, so we need to be more clever in proving linear independence. Ways to be more clever include:

1. Choosing three different times t_1, t_2, t_3 at which to sample equation (6.10).

2. Multiplying (6.10) by some other function of t and then taking a limit as $t \to \infty$ (or taking a limit as $t \to t^*$ for some particular $t^* \in \mathbb{R}$)
3. Differentiating equation (6.10) (if the left-hand-side is differentiable).

Claim: $\{\cos(t), t, t^2\}$ are linearly independent functions in V.

Proof: Suppose there are scalars $a_1, a_2, a_3 \in \mathbb{R}$ such that

$$a_1 \cos(t) + a_2 t + a_3 t^2 = 0 \quad \forall t \in \mathbb{R} \tag{6.11}$$

We want to show $a_1 = a_2 = a_3 = 0$. Sampling (6.11) at time $t = 0$ gives

$$a_1 \cos(0) + 0 + 0 = 0$$

and since $\cos(0) = 1$, this means

$$\boxed{a_1 = 0}$$

Substituting $a_1 = 0$ back into (6.11) gives

$$a_2 t + a_3 t^2 = 0 \quad \forall t \in \mathbb{R} \tag{6.12}$$

Differentiating both sides gives

$$a_2 + 2a_3 t = 0 \quad \forall t \in \mathbb{R} \tag{6.13}$$

Sampling (6.13) at time $t = 0$ gives

$$\boxed{a_2 = 0}$$

Substituting $a_2 = 0$ back into (6.12) gives

$$a_3 t^2 = 0 \quad \forall t \in \mathbb{R}$$

Sampling the above equation at $t = 1$ gives

$$a_3 \cdot 1^2 = 0 \implies \boxed{a_3 = 0}$$

so we have proven $a_1 = a_2 = a_3 = 0$.

6.8.2. Linear independence of $\{\cos(t+5), (t+1)^{12}\}$

Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$.

Claim: $\{\cos(t+5), (t+1)^{12}\}$ are linearly independent functions in V.

Proof: Suppose a and b are some real numbers such that

$$a\cos(t+5) + b(t+1)^{12} = 0 \quad \forall t \in \mathbb{R} \tag{6.14}$$

We want to show $a = b = 0$. Fixing $t > 0$ and dividing both sides of (6.14) by $(t+1)^{12}$ gives

$$\frac{a\cos(t+5)}{(t+1)^{12}} + b = 0 \quad \forall t > 0$$

Thus

$$\underbrace{\lim_{t\to\infty}\left(\frac{a\cos(t+5)}{(t+1)^{12}} + b\right)}_{b} = \underbrace{\lim_{t\to\infty} 0}_{0}$$

So $b = 0$. Substituting $b = 0$ back into (6.14) gives

$$a\cos(t+5) = 0 \quad \forall t \in \mathbb{R}$$

Sampling this equation at time $t = -5$ gives $a\cos(-5+5) = 0$. We conclude $a = 0$.

6.8.3. Basis for polynomials of degree ≤ 3

Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$. Define the subspace

$$V_3 = \{f \in V : f \text{ is a polynomial of degree} \leq 3\}$$

We claim that $\{1, t, t^2, t^3\}$ is a basis for V_3 (and so $\dim(V_3) = 4$).

- Proof that $\{1, t, t^2, t^3\}$ is linearly independent: Suppose

  $$a + bt + ct^2 + dt^3 = 0 \quad \forall t \in \mathbb{R} \tag{6.15}$$

  for some $a, b, c, d \in \mathbb{R}$. We want to show $a = b = c = d = 0$. By repeated differentiation of equation (6.15) we obtain

  $$b + 2ct + 3dt^2 = 0 \quad \forall t \in \mathbb{R} \tag{6.16}$$

  $$2c + 6dt = 0 \quad \forall t \in \mathbb{R} \tag{6.17}$$

  $$6d = 0 \tag{6.18}$$

  The equation $6d = 0$ implies $d = 0$. Substituting $d = 0$ into (6.17) gives $2c = 0$, which implies $c = 0$. Substituting $d = c = 0$ into (6.16) gives $b = 0$ and we are done.

- Proof that $Span\{1, t, t^2, t^3\} = V_3$: First observe that the functions $1, t, t^2, t^3$ are themselves vectors in the subspace V_3 (so any linear combination of these functions is again

in V_3). To show that every function in V_3 is in $Span\{1, t, t^2, t^3\}$, fix $f \in V_3$. Then

$$f(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \quad \forall t \in \mathbb{R}$$

for some $a_0, a_1, a_2, a_3 \in \mathbb{R}$. So

$$f = a_0 + a_1 t + a_2 t^2 + a_3 t^3 \in Span\{1, t, t^2, t^3\}$$

The basis $\{1, t, t^2, t^3\}$ is not the only basis for V_3. The reader should verify $\{2, t, t^2, t^3\}$ and $\{1 + t, t, t^2, t^3 - 4t^2\}$ are also valid bases for V_3. Of course, all bases for V_3 contain exactly four vectors because $\dim(V_3) = 4$.

6.8.4. A basis for polynomial functions of degree at most k

Let V be the real vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$. So $f \in V$ implies

$$f(t) = a_0 + a_1 t + a_2 t^2 + \cdots + a_k t^k \quad \forall t \in \mathbb{R}$$

for some $k \in \mathbb{N}$ and some real numbers $a_0, a_1, \ldots, a_k$. We say the polynomial has *degree* k if $a_k \neq 0$ (where k is a positive integer and a_k is the coefficient of the largest power of t). We say the polynomial has *degree* 0 if $f(t) = a_0$ for all $t \in \mathbb{R}$ (so it is a constant function).

LEMMA 6.5. *For all $k \in \{0, 1, 2, \ldots\}$, the list of $k + 1$ functions $\{1, t, t^2, \ldots, t^k\}$ is linearly independent. (By the case $k = 0$, we mean $\{1\}$, the list that has just the always-one function.)*

PROOF. We use induction on k. The list $\{1\}$ is linearly independent since $1 \neq 0$. Now suppose the list $\{1, t, \ldots, t^k\}$ is linearly independent for some $k \in \{0, 1, 2, \ldots\}$ (it holds for $k = 0$). We now prove the list $\{1, t, \ldots, t^k, t^{k+1}\}$ is linearly independent. Suppose

$$a_0 \cdot 1 + a_1 t + a_2 t^2 + \cdots + a_k t^k + a_{k+1} t^{k+1} = 0 \quad \forall t \in \mathbb{R} \tag{6.19}$$

We want to show $a_i = 0$ for all $i \in \{0, 1, 2, \ldots, k + 1\}$. Fixing $t > 0$ and dividing by t^{k+1} gives

$$\frac{a_0 \cdot 1 + a_1 t + a_2 t^2 + \cdots + a_k t^k}{t^{k+1}} + a_{k+1} = 0 \quad \forall t > 0$$

Taking $t \to \infty$ gives $a_{k+1} = 0$. Substituting $a_{k+1} = 0$ into equation (6.19) gives

$$a_0 \cdot 1 + a_1 t + a_2 t^2 + \cdots + a_k t^k = 0 \quad \forall t \in \mathbb{R}$$

so $a_0 = a_1 = \cdots = a_k = 0$ (since we have assumed $\{1, t, \ldots, t^k\}$ is linearly independent). □

Fix $k \in \mathbb{N}$. Define $V_k \subseteq V$ by

$$\begin{aligned} V_k &= \{f \in V : f \text{ has degree} \leq k\} \\ &= \{f \in V : f(t) = a_0 + a_1 t + \cdots + a_k t^k \quad \text{for some } a_i \in \mathbb{R}\} \end{aligned}$$

The set V_k is a subspace because $0 \in V_k$, and the linear combination of two polynomials of degree at most k is again a polynomial of degree at most k. The functions $1, t, t^2, ..., t^k$ are all in V_k, and any other function in V_k is a linear combination of these. So

$$V_k = Span\{1, t, t^2, ..., t^k\}$$

The above lemma ensures the list $\{1, t, ..., t^k\}$ is linearly independent, so $\{1, t, ..., t^k\}$ is a basis for V_k and $\dim(V_k) = k + 1$. Exercise 6.59 shows that any list of nonzero polynomials that have different degrees is linearly independent.

6.8.5. A basis for $\{f \in V_3 : f(0) = f(1)\}$

Let V_3 be the real vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$ of degree less than or equal to 3. Define

$$S = \{f \in V_3 : f(0) = f(1)\}$$

Claim: S is a subspace.

Proof: Clearly the all-zero function is in S. Now fix $f, g \in S$ and $a, b \in \mathbb{R}$. We want to show $af + bg \in S$. We have

$$\begin{aligned} (af + bg)(0) &= af(0) + bg(0) \\ &\overset{(a)}{=} af(1) + bg(1) \\ &= (af + bg)(1) \end{aligned}$$

where (a) holds because $f \in S$ and $g \in S$. Thus, $af + bg \in S$. □

We know S is not the trivial subspace because $1 \in S$. Also, $S \subseteq V_3$ but S does not include *everything* in V_3 (for example, $t \notin S$ but $t \in V_3$). Thus

$$\{0\} \subsetneq S \subsetneq V_3$$

and so (by Lemma 6.4):

$$0 < \dim(S) < \dim(V_3) = 4$$

So $\dim(S) \in \{1, 2, 3\}$. Since S is nontrivial and finite dimensional, it must have a basis

(Lemma 6.2). Observe that:

$$1 \in S$$
$$t(t-1) \in S$$
$$t^2(t-1) \in S$$

The reader can verify $\{1, t(t-1), t^2(t-1)\}$ are linearly independent in S (see Exercise 6.47). So the list can be increased to a basis for S. But there are already three vectors in this list, and S has dimension at most 3, so $\{1, t(t-1), t^2(t-1)\}$ is a basis for S (and $\dim(S) = 3$).

6.8.6. Subspaces of infinite sequences

Let $\mathbb{R}^{\mathbb{N}}$ denote the real vector space of infinite sequences $(x_1, x_2, x_3, \ldots)$ where $x_i \in \mathbb{R}$ for $i \in \mathbb{N}$. Define $S \subseteq \mathbb{R}^{\mathbb{N}}$ by

$$S = \left\{(x_1, x_2, x_3, \ldots) \in \mathbb{R}^{\mathbb{N}} : x_{k+2} + 3x_{k+1} + 2x_k = 0 \quad \forall k \in \mathbb{N}\right\}$$

Claim:

1. S is a subspace of $\mathbb{R}^{\mathbb{N}}$.
2. A vector $(x_1, x_2, x_3, \ldots) \in S$ is completely determined by knowing (x_1, x_2).
3. $\dim(S) = 2$, and a basis for S is $\{(u_k), (v_k)\}$, where

$$(u_k) = ((-1)^k)_{k=1}^{\infty} = (-1, 1, -1, 1, \ldots)$$
$$(v_k) = ((-2)^k)_{k=1}^{\infty} = (-2, 4, -8, 16, \ldots)$$

That is, for every $(x_1, x_2, x_3, \ldots) \in S$, there are scalars $a, b \in \mathbb{R}$ such that

$$x_k = a(-1)^k + b(-2)^k \quad \forall k \in \mathbb{N}$$

Proof:

1. To show S is a subspace, first note that the always-zero sequence is in S. Now let (x_k) and (y_k) be two vectors in S. Fix $a, b \in \mathbb{R}$. Define

$$(h_k) = a(x_k) + b(y_k)$$

so $h_k = ax_k + by_k$ for all $k \in \mathbb{N}$. We want to show $(h_k) \in S$. For each $k \in \mathbb{N}$ we have

$$\begin{aligned} h_{k+2} + 3h_{k+1} + 2h_k &= (ax_{k+2} + by_{k+2}) + 3(ax_{k+1} + by_{k+1}) + 2(ax_k + by_k) \\ &= a(x_{k+2} + 3x_{k+1} + 2x_k) + b(y_{k+2} + 3y_{k+1} + 2y_k) \end{aligned}$$

$$\overset{(a)}{=} a0 + b0 = 0$$

where (a) holds because $(x_k) \in S$ and $(y_k) \in S$. Thus, $(h_k) \in S$.

2. If we know x_1, x_2, we obtain $x_3 = -3x_2 - 2x_1$. In general we obtain $x_k = -3x_{k-1} - 2x_{k-2}$ for all $k \geq 3$.

3. The equation $x_{k+2} + 3x_{k+1} + 2x_k = 0$ for $k \in \mathbb{N}$ is called a *linear difference equation*. We can search for nontrivial solutions by considering sequences of the type $x_k = \theta^k$ for some scalar $\theta \neq 0$. For such a sequence to be in S we need

$$\theta^{k+2} + 3\theta^{k+1} + 2\theta^k = 0 \quad \forall k \in \{1, 2, 3, \ldots\}$$

It suffices to choose θ to satisfy

$$\theta^2 + 3\theta + 2 = 0 \iff (\theta + 1)(\theta + 2) = 0$$

So we use $\theta_1 = -1$ and $\theta_2 = -2$. Thus, two distinct sequences in S are

$$(u_k) = ((-1)^k)_{k=1}^{\infty} = (-1, 1, -1, 1, \ldots)$$
$$(v_k) = ((-2)^k)_{k=1}^{\infty} = (-2, 4, -8, 16, \ldots)$$

Since S is a subspace, it follows that $a(u_k) + b(v_k) \in S$ for all $a, b \in \mathbb{R}$.

We now show $S = Span\{(u_k), (v_k)\}$, so *every* sequence $(x_k) \in S$ is a linear combination of (u_k) and (v_k). Fix $(x_k) \in S$. We want to show $x_k \in Span\{(u_k), (v_k)\}$. Since $(x_k) = (x_1, x_2, x_3, \ldots)$, bullet 2 implies (x_k) is the *only* sequence in S that starts out with the first two values (x_1, x_2). On the other hand, we can always find $(a, b) \in \mathbb{R}$ to solve

$$a(-1, 1) + b(-2, 4) = (x_1, x_2)$$

This is because $\{(-1, 1), (-2, 4)\}$ are linearly independent in $\mathbb{R}^2$, so they must span $\mathbb{R}^2$. Thus, the sequence $(h_k) = a(u_k) + b(v_k)$ is another sequence in S that has first two values (x_1, x_2). But (x_k) is the only such sequence, and so $(h_k) = (x_k)$. That is, $(x_k) = a(u_k) + b(v_k)$. This completes the proof that $S = Span\{(u_k), (v_k)\}$.

To show $\{(u_k), (v_k)\}$ are linearly independent, suppose

$$a(u_k) + b(v_k) = 0$$

for some $a, b \in \mathbb{R}$. We want to show $a = b = 0$. We know

$$a(-1, 1, -1, \ldots) + b(-2, 4, -8, \ldots) = (0, 0, 0, \ldots)$$

Just by looking at the first two entries in the above equation we have

$$\begin{aligned} a(-1) + b(-2) &= 0 \\ a(1) + b(4) &= 0 \end{aligned}$$

from which it easily follows that $a = b = 0$.

6.8.7. Complex sequences

Let $\mathbb{R}^{\mathbb{N}}$ and $\mathbb{C}^{\mathbb{N}}$ respectively denote the real vector space of real-valued sequences, and the complex vector space of complex-valued sequences. Define $S \subseteq \mathbb{R}^{\mathbb{N}}$ and $H \subseteq \mathbb{C}^{\mathbb{N}}$ by

$$\begin{aligned} S &= \{(x_1, x_2, x_3, \ldots) \in \mathbb{R}^{\mathbb{N}} : x_{k+2} + x_{k+1} + x_k = 0 \quad \forall k \in \mathbb{N}\} \\ H &= \{(z_1, z_2, z_3, \ldots) \in \mathbb{C}^{\mathbb{N}} : z_{k+2} + z_{k+1} + z_k = 0 \quad \forall k \in \mathbb{N}\} \end{aligned}$$

It is not difficult to show that S and H are indeed subspaces of their respective vector spaces $\mathbb{R}^{\mathbb{N}}$ and $\mathbb{C}^{\mathbb{N}}$. Here we show $\dim(S) = 2$ and $\dim(H) = 2$.

It turns out that it is easier to obtain a basis for H, so we do that first: To solve the linear difference equation $z_{k+2} + z_{k+1} + z_k = 0$, we search for solutions of the form $z_k = \theta^k$ for some (possibly complex-valued) scalar $\theta \neq 0$. Such a solution would satisfy

$$\theta^{k+2} + \theta^{k+1} + \theta^k = 0 \quad \forall k \in \mathbb{N}$$

It suffices to find a scalar θ that satisfies $\theta^2 + \theta + 1 = 0$. There are two such θ values:

$$\begin{aligned} \theta_1 &= \frac{-1}{2} + i\frac{\sqrt{3}}{2} \\ \theta_2 &= \frac{-1}{2} - i\frac{\sqrt{3}}{2} \end{aligned}$$

So $(u_k) = (\theta_1^k)$ and $(v_k) = (\theta_2^k)$ are two distinct sequences in the subspace $H \subseteq \mathbb{C}^{\mathbb{N}}$. By an argument similar to that of the previous subsection, it can be shown that $\{(u_k), (v_k)\}$ is a basis for H, so $\dim(H) = 2$.

What about a basis for S? The sequences $\{(u_k), (v_k)\}$ do not work because they are not real-valued sequences. However, we know that $(u_k) \in H$ and so

$$u_{k+2} + u_{k+1} + u_k = 0 \quad \forall k \in \mathbb{N}$$

Taking the *complex conjugate* of both sides gives:[2]

$$\overline{u}_{k+2} + \overline{u}_{k+1} + \overline{u}_k = 0 \quad \forall k \in \mathbb{N}$$

where we have used the fact that the complex conjugate of a sum is the sum of complex conjugates. Thus, the sequence $(\overline{u}_k)$, defined by taking complex conjugates of all entries in the sequence (u_k), is also in H. Since H is a subspace, any linear combination of (u_k) and $(\overline{u}_k)$ must be in H. Define

$$(r_k) = \frac{1}{2}(u_k) + \frac{1}{2}(\overline{u}_k) \in H$$
$$(s_k) = \frac{1}{2i}(u_k) - \frac{1}{2i}(\overline{u}_k) \in H$$

so $r_k = Re(u_k)$ and $s_k = Im(u_k)$ for all $k \in \mathbb{N}$. Then (r_k) and (s_k) are *real-valued sequences* that satisfy the desired linear difference equation, so they are indeed in S. Using an argument similar to the previous subsection, it can be shown that $\{(r_k), (s_k)\}$ is a basis for S. This method of working with real-valued linear difference equations by first allowing solutions to be complex-valued, then taking real and imaginary parts, is a useful technique.

6.9. Linear transformations

Let U, V be two vector spaces over the same field $\mathbb{F}$. A function $T : U \to V$ is called a *linear transformation* if

$$T(ax + by) = aT(x) + bT(y)$$

for all scalars $a, b \in \mathbb{F}$ and all vectors $x, y \in U$. A linear transformation with $V = U$ is called a *linear operator*.

6.9.1. Equivalent definition

A function $T : U \to V$ is a linear transformation if and only if the following two properties hold:

$$T(u + w) = T(u) + T(w) \quad \forall u, w \in U \tag{6.20}$$

$$T(au) = aT(u) \quad \forall u \in U, a \in \mathbb{F} \tag{6.21}$$

[2] For a complex number $z = a + bi$, where $a, b \in \mathbb{R}$, we define $Re(z) = a$ (the real part of z) and $Im(z) = b$ (the imaginary part of z). Also, the complex conjugate $\overline{z}$ is defined $\overline{z} = a - bi$.

These two properties can be viewed as an equivalent definition of a linear transformation. If $T : U \to V$ satisfies the original definition then for any $u, w \in U$ and $a \in \mathbb{F}$ we have

$$T(u+w) = T(1 \cdot u + 1 \cdot w) = 1T(u) + 1T(w) = T(u) + T(w)$$
$$T(au) = T(au + 0u) = aT(u) + 0T(u) = aT(u)$$

so T also satisfies properties (6.20)-(6.21). Conversely, if T satisfies properties (6.20)-(6.21), the reader can verify that T is indeed a linear transformation.

In general, both properties (6.20)-(6.21) are needed to ensure a function $T : U \to V$ is a linear transformation. Notes:

- Exercise 6.53 treats an example function $T : \mathbb{R}^2 \to \mathbb{R}^2$ that satisfies (6.21) but not (6.20).
- Exercise 6.54 treats an example function $T : \mathbb{C} \to \mathbb{C}$ that satisfies (6.20) but not (6.21).
- For certain special case fields $\mathbb{F}$, the property (6.20) implies property (6.21). This is true for: the field $\mathbb{Q}$ of rational numbers (Exercise 6.55); the finite field $\mathbb{F}_p = \{0, 1, \ldots, p-1\}$ with mod p arithmetic, where p is prime (Exercise 6.56). [It does not necessarily hold for finite fields of non-prime sizes, see Exercise 6.58.]
- For the field $\mathbb{F} = \mathbb{R}$, any function $T : U \to V$ that satisfies property (6.20) must also satisfy $T(au) = aT(u)$ for all $u \in U$ and all *rational numbers* $a \in \mathbb{Q}$ (Exercise 6.55). If we additionally assume $U = \mathbb{R}^n$, $V = \mathbb{R}^m$, and if we assume function T is continuous at 0, then property (6.21) must also extend to irrational numbers a. However, a set theory axiom call the *axiom of choice* can be used to prove existence of a discontinuous function $T : \mathbb{R}^n \to \mathbb{R}^m$ that satisfies property (6.20) but not property (6.21) (Exercise 6.57). [Note that "continuity" of a general map $T : U \to V$ cannot be defined unless the vector spaces U, V have additional structure, such as the *inner product structure* of Chapter 8 that allows distances between vectors to be defined.]

6.9.2. Basic properties

Suppose $T : U \to V$ is a linear transformation. For any vectors $x, y, z \in U$ we have

$$T(x+y+z) = T((x+y)+z) \overset{(a)}{=} T(x+y) + T(z) = T(x) + T(y) + T(z)$$

where (a) uses (6.20) with $u = x + y$ and $w = z$. This uses a known property about evaluating T at the sum of *two* vectors to prove a property about evaluating T at the sum of *three* vectors. More generally, for any $k \in \mathbb{N}$ and any $x_1, \ldots, x_k \in U$, $a_1, \ldots, a_k \in \mathbb{F}$:

$$\boxed{T\left(\sum_{i=1}^{k} a_i x_i\right) = \sum_{i=1}^{k} a_i T(x_i)}$$

A linear transformation $T : U \to V$ always satisfies

$$\boxed{T(0) = 0}$$

That is, T maps the vector $0 \in U$ to the vector $0 \in V$. This is because

$$T(0) = T(0 + 0) = T(0) + T(0)$$

Since $T(0) \in V$, Axiom 4 for the vector space V ensures existence of $-T(0)$. Adding $-T(0)$ to both sides of the equation $T(0) = T(0) + T(0)$ gives $0 = T(0)$.

6.10. $Null(T)$ and $Span(T)$

Let $T : U \to V$ be a linear transformation. Define the *null space of T*, denoted $Null(T)$, by

$$Null(T) = \{x \in U : T(x) = 0\}$$

Define $Span(T)$, called the span of T, by

$$Span(T) = \{T(x) \in V : x \in U\}$$

It can be shown that $Null(T)$ is a subspace of U and $Span(T)$ is a subspace of V (see Exercise 6.4). The *rank of T*, denoted rank(T), is defined as the dimension of $Span(T)$:

$$\text{rank}(T) = \dim(Span(T))$$

Equations of the form $T(x) = b$ are similar to matrix equations $Ax = b$. Fix $b \in V$ and suppose x^* is a particular vector in U such that $T(x^*) = b$. The reader can verify that

$$\{x \in U : T(x) = b\} = \{x^* + v \in U : v \in Null(T)\}$$

To prove this, show: (i) If x is a vector in U that solves $T(x) = b$, then $x = x^* + v$ for some $v \in Null(T)$; (ii) If $v \in Null(T)$ then $x^* + v$ solves the equation $T(x) = b$.

6.10.1. Terminology

- $Null(T)$ is also called the *kernel* of T, denoted $ker(T)$.
- $Span(T)$ is also called $Range(T)$ or $Image(T)$.
- For $T : U \to V$, $Span(T)$ can also be denoted $T(U)$.

6.11. Example linear transformations

1. Fix $\mathbb{F}$ as a field. Fix $m, n \in \mathbb{N}$. Let A be a $m \times n$ matrix with entries in $\mathbb{F}$. The function $T : \mathbb{F}^n \to \mathbb{F}^m$ defined by $T(x) = Ax$ for all $x \in \mathbb{F}^n$ is a linear transformation. Indeed, for any $x, y \in \mathbb{F}^n$ and $a, b \in \mathbb{F}$ we have

$$\begin{aligned} T(ax + by) &= A(ax + by) \\ &= a(Ax) + b(Ay) \\ &= aT(x) + bT(y) \end{aligned}$$

   Also, $Null(T) = Null(A)$ and $Span(T) = Col(A)$.

2. Let U be the real vector space of differentiable functions $f : \mathbb{R} \to \mathbb{R}$. Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$ (not necessarily differentiable). Define $T : U \to V$ by

$$T(f) = f' \quad \forall f \in V \tag{6.22}$$

   So T maps each differentiable function in U to its derivative function in V. We use $V \neq U$ because the derivative function is not necessarily differentiable. To show that T is a linear transformation, let f, g be functions in U and fix $a, b \in \mathbb{R}$. Then

$$T(af + bg) = (af + bg)' = af' + bg' = aT(f) + bT(g)$$

   $Null(T)$ is the set of all constant functions (since constant functions map to the always zero function). $Span(T)$ is the set of all possible derivative functions (we do not attempt to characterize them further).

   The following are input-output examples for this transformation T. The input is always a function in U, and the output is always a function in V (the last example assumes h is some function in U):

$$T(e^{-2t}) = -2e^{-2t} \tag{6.23}$$

$$T(8t^2 - t + 4) = 16t - 1 \tag{6.24}$$

$$T(7t) = 7 \tag{6.25}$$

$$T(3) = 0 \tag{6.26}$$

$$T(h(t-3)) = h'(t-3) \tag{6.27}$$

   with the understanding that e^{-2t}, $8t^2 - t + 4$, $7t$, 3, $h(t-3)$, which are the arguments of $T(\cdot)$ on the left-hand-sides of (6.23)-(6.27), represent differentiable functions over the

entire timeline. We have assumed h is differentiable and so its shifted function $h(t-3)$ is differentiable. Also, $-2e^{-2t}$, $16t-1$, 1, 0, $h'(t-3)$, which are the right-hand-sides of (6.23)-(6.27), represent functions over the *entire timeline*. If we want to evaluate the output functions at a particular time, say $t=2$, we have

$$T(e^{-2t})(2) = -2e^{-4} \tag{6.28}$$

$$T(8t^2 - t + 4)(2) = 31 \tag{6.29}$$

$$T(7t)(2) = 7 \tag{6.30}$$

$$T(3)(2) = 0 \tag{6.31}$$

$$T(h(t-3))(2) = h'(-1) \tag{6.32}$$

where the right-hand-sides and left-hand-sides of (6.28)-(6.32) are just real numbers.

An important notational issue: When evaluating the function in (6.23) at $t=2$, it would not be correct to write $T(e^{-4}) = -2e^{-4}$, since the input to T needs to be a *function*, not a real number. Therefore, $T(e^{-4})$ only makes sense if e^{-4} is viewed as a constant function that is always e^{-4}, in which case $T(e^{-4}) = 0$ because the derivative of a constant function is the zero function.
To avoid ambiguity about whether we mean e^{-2t} to be a function over all time, or a real number that is evaluated at some particular time t, we could represent equation (6.23) as follows: Define $s : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ by

$$s(t) = e^{-2t} \quad \forall t \in \mathbb{R}$$
$$g(t) = -2e^{-2t} \quad \forall t \in \mathbb{R}$$

Then (6.23) can be expressed:

$$T(s) = g$$

It is common to represent a function $f : \mathbb{R} \to \mathbb{R}$ as either f or $f(t)$. Using f emphasizes that we are talking about the entire function that maps $\mathbb{R} \to \mathbb{R}$. Using $f(t)$ leaves an ambiguity: Do we mean the entire function, or do we mean a real number that evaluates a function at a particular time? The ambiguity is typically resolved by the context. For the same reason, it is acceptable for students to use the following alternative notation to introduce the transformation $T(f) = f'$ in (6.22):

- Notational variation 1: $T(f) = f'(t)$
- Notational variation 2: $T(f)(t) = f'(t) \quad \forall t \in \mathbb{R}$
- Notational variation 3: $T(f) = \left(f'(t), t \in \mathbb{R}\right)$

3. Let V be the real vector space of all functions $f : \mathbb{R} \to \mathbb{R}$. Define $T : V \to V$ such that an input function $f \in V$ yields output function $T(f)$ defined by

$$T(f)(t) = e^t f(t) \quad \forall t \in \mathbb{R} \tag{6.33}$$

So T takes a function $f \in V$ and multiplies it by the exponential function. To show T is a linear transformation, fix $f, g \in V$ and $a, b \in \mathbb{R}$. For all $t \in \mathbb{R}$

$$\begin{aligned} T(af + bg)(t) &= e^t(af(t) + bg(t)) \\ &= ae^t f(t) + be^t g(t) \\ &= aT(f)(t) + bT(g)(t) \\ &= (aT(f) + bT(g))(t) \end{aligned}$$

where the last equality uses (6.6). This holds for all $t \in \mathbb{R}$, so $T(af + bg) = aT(f) + bT(g)$ (recall (6.5)). $Null(T) = \{0\}$ because the only function $f \in V$ that satisfies $e^t f(t) = 0$ for all t is the always-zero function. $Span(T) = V$ because for any $g \in V$ we can define $f \in V$ by $f(t) = g(t)e^{-t}$ for all $t \in \mathbb{R}$, so $T(f) = g$. A similar problem is in Exercise 6.10.

Acceptable notation: In view of the discussion in the last paragraph of the previous bullet, it is acceptable for students to introduce the transformation T in (6.33) by

$$T(f) = e^t f(t)$$

where the context makes it clear that $e^t f(t)$ is viewed as a function over the *entire timeline*. Similarly, to prove T is a linear transformation it is acceptable to write:

$$\begin{aligned} T(af + bg) &= e^t(af(t) + bg(t)) \\ &= ae^t f(t) + be^t g(t) \\ &= aT(f) + bT(g) \end{aligned}$$

Another way to introduce the function T in (6.33) is

$$T(f) = \left(e^t f(t), t \in \mathbb{R}\right) \quad \forall f \in V$$

4. Consider the vector space $\mathbb{F}^{\mathbb{N}}$ of infinite sequences $(x_1, x_2, x_3, \ldots)$ with entries in some field $\mathbb{F}$. Define $T : \mathbb{F}^{\mathbb{N}} \to \mathbb{F}^{\mathbb{N}}$ by a "left-shift-by-one":

$$T((x_1, x_2, x_3, \ldots)) = (x_2, x_3, x_4, \ldots) \quad \forall x \in \mathbb{F}^{\mathbb{N}}$$

To show T is a linear transformation, fix $x, y \in \mathbb{F}^{\mathbb{N}}$ and $a, b \in \mathbb{F}$. So $x = (x_1, x_2, \ldots)$,

$y = (y_1, y_2, ...)$ and

$$\begin{aligned} T(ax + by) &= T((ax_1 + by_1, ax_2 + by_2, ax_3 + by_3, ...)) \\ &= (ax_2 + by_2, ax_3 + by_3, ...) \\ &= a(x_2, x_3, ...) + b(y_2, y_3, ...) \\ &= aT(x) + bT(y) \end{aligned}$$

The subspaces $Span(T)$ and $Null(T)$ are explored in Exercise 6.51.

5. Let $(\Omega, \mathcal{F}, P)$ be a probability space. Define V as the real vector space of all random variables $X : \Omega \to \mathbb{R}$ that have finite expectation. Define $T : V \to \mathbb{R}$ by

$$T(X) = \mathbb{E}[X] \quad \forall X \in V$$

The fact that T is a linear transformation follows by *linearity of expectation*. Also, $Null(T) = \{X \in V : \mathbb{E}[X] = 0\}$ and $Span(T) = \mathbb{R}$ (consider constant random variables).

The linear transformations in examples 3 and 4 above are linear operators because they have the form $T : V \to V$.

6.12. Linear transformations are defined by their values on a basis

If $\dim(U) < \infty$, then a linear transformation $T : U \to V$ can be described just by knowing its values on any basis of U. Specifically, suppose $\dim(U) = n \in \mathbb{N}$. Let $\{x_1, ..., x_n\}$ be any basis of U. Sample T on this basis and define $y_1, ..., y_n$ as the resulting output:

$$y_i = T(x_i) \quad \forall i \in \{1, \ldots, n\}$$

Given any $x \in U$, let $(a_1, ..., a_n)$ be the unique scalars in $\mathbb{F}$ that satisfy

$$x = \sum_{i=1}^{n} a_i x_i$$

Such scalars exist and are unique because $\{x_1, ..., x_n\}$ is a basis for U. Then we have

$$T(x) = T\left(\sum_{i=1}^{n} a_i x_i\right) = \sum_{i=1}^{n} a_i T(x_i) = \sum_{i=1}^{n} a_i y_i$$

This means that for any $x \in U$, the value of $T(x)$ can be computed by first finding $(a_1, ..., a_n)$ (which are found only from x, without using any information about T) and using them as weights to linearly combine the sampled output vectors $y_1, ..., y_n$.

6.13. Invertible linear transformations

Suppose U and V are nonempty sets. A function $T : U \to V$ is *invertible* if there is a function $T^{-1} : V \to U$ that satisfies:

$$T^{-1}(T(u)) = u \quad \forall u \in U$$
$$T(T^{-1}(v)) = v \quad \forall v \in V$$

The function T^{-1} is called the *inverse function* of T. A function $T : U \to V$ that is invertible is also called a *bijection* or a *one-to-one correspondence* between the sets U and V. If a function is invertible, it has a unique inverse function, the inverse function is invertible, and the inverse of the inverse is the original function.

Suppose U and V are vector spaces over some field $\mathbb{F}$. Suppose $T : U \to V$ is an invertible linear transformation. It can be shown that:

- $T^{-1}(0) = 0$.
- The inverse function $T^{-1} : V \to U$ is also an invertible linear transformation.
- $Null(T) = \{0\}$ and $\text{rank}(T) = \dim(U) = \dim(V)$. [Possibly $\dim(U) = \dim(V) = \infty$]
- $\{u_1, ..., u_k\}$ is a basis for U if and only if $\{T(u_1), ..., T(u_k)\}$ is a basis for V.
- Let S be a subspace of U. Define $T(S) = \{T(s) : s \in S\}$. Then $\dim(S) = \dim(T(S))$.

For example, let U be the real vector space of polynomial functions $p : \mathbb{R} \to \mathbb{R}$ of degree less than or equal to 3. The function $T : U \to \mathbb{R}^4$ that maps each polynomial $a_0 + a_1 t + a_2 t^2 + a_3 t^3 \in U$ to the vector $(a_0, a_1, a_2, a_3) \in \mathbb{R}^4$ is an invertible linear transformation. So $\text{rank}(T) = \dim(U) = \dim(\mathbb{R}^4) = 4$. The dimension of the subspace $S = \{p \in U : p(1) = 0, p(2) = 0\}$ is the same as the dimension of its corresponding subspace in $\mathbb{R}^4$.

6.14. The matrix of a linear transformation

Let U and V be nontrivial vector spaces over a common field $\mathbb{F}$. If U and V are both finite dimensional then a linear transformation $T : U \to V$ can be equivalently described by a matrix. To see this, suppose $\dim(U) = n$ and $\dim(V) = m$ for some $n, m \in \mathbb{N}$. Let $\{u_1, ..., u_n\}$ and $\{v_1, ..., v_m\}$ be bases for U and V, respectively. For each $j \in \{1, ..., n\}$, define $a_{ij} \in \mathbb{F}$ for $i \in \{1, ..., m\}$ according to the unique basis representation of $T(u_j) \in V$:

$$T(u_j) = \sum_{i=1}^{m} a_{ij} v_i \quad \forall j \in \{1, ..., n\} \tag{6.34}$$

Define the $m \times n$ matrix $A = (a_{ij})$. All coefficients of A are scalars in the field $\mathbb{F}$.

Fix $x \in U$ and define $y = T(x)$. Since $x \in U$ and $y \in V$, we can write x and y according to their unique linear combinations using the basis vectors $\{u_1, ..., u_n\}$ and $\{v_1, ..., v_m\}$:

$$x = \sum_{j=1}^{n} c_j u_j \tag{6.35}$$

$$y = \sum_{i=1}^{m} \theta_i v_i \tag{6.36}$$

for some unique scalars c_j and θ_i. Define $c = (c_1, \ldots, c_n)$ and $\theta = (\theta_1, \ldots, \theta_m)$. We now show

$$\theta = Ac \tag{6.37}$$

To prove (6.37), we have

$$y = T(x) \overset{(a)}{=} T\left(\sum_{j=1}^{n} c_j u_j\right) = \sum_{j=1}^{n} c_j T(u_j) \overset{(b)}{=} \sum_{j=1}^{n} c_j \sum_{i=1}^{m} a_{ij} v_i \overset{(c)}{=} \sum_{i=1}^{m} \underbrace{\left(\sum_{j=1}^{n} c_j a_{ij}\right)}_{i\text{th row of } Ac} v_i$$

where (a) holds by (6.35); (b) holds by (6.34); (c) holds by switching the sums and using the definition of the ith row of Ac. The above expresses y as a linear combination of the basis vectors $\{v_1, ..., v_m\}$, but there is only one way to do this, so (6.36) implies that for all $i \in \{1, ..., m\}$

$$\theta_i = \sum_{j=1}^{n} c_j a_{ij} = i\text{th row of } Ac$$

which means $\theta = Ac$, proving (6.37). In summary, the input-output relationship $y = T(x)$ for the linear transformation T between the general (finite dimensional) vector spaces U and V can be expressed as a mundane matrix multiplication $\theta = Ac$ for $c \in \mathbb{F}^n$ and $\theta \in \mathbb{F}^m$, where c and θ uniquely relate to x and y via (6.35) and (6.36). We say that A is the *matrix of* T, written $Mat(T)$, associated with the bases $\{u_1, ..., u_n\}$ and $\{v_1, ..., v_m\}$ for U and V.

For example, let V be the real vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$ with degree less than or equal to 2. Let $T : V \to V$ be the linear transformation that maps each $f \in V$ to $T(f) = f'$. Let $\{1, t, t^2\}$ be the basis for both input and output sides of T. We can write

$$a + bt + ct^2 \equiv \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

where the left-hand-side expresses the polynomial in V, and the right-hand-side equiva-

lently represents the polynomial by its coefficients under the basis $\{1, t, t^2\}$. Then

$$\left(a + bt + ct^2 \overset{T}{\to} b + 2ct\right) \iff \begin{bmatrix} a \\ b \\ c \end{bmatrix} \overset{A}{\to} \begin{bmatrix} b \\ 2c \\ 0 \end{bmatrix} \quad , \quad A = Mat(T) = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 2 \\ 0 & 0 & 0 \end{bmatrix}$$

In particular, each column of $Mat(T)$ corresponds to evaluating T on the corresponding basis vector:

$$T(1) = 0 \equiv \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix} \quad , \quad T(t) = 1 \equiv \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad , \quad T(t^2) = 2t \equiv \begin{bmatrix} 0 \\ 2 \\ 0 \end{bmatrix}$$

For another example, let V be the same real vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$ of degree less than or equal to 2. Use the same basis $\{1, t, t^2\}$ for this space. Consider the function $G : V \to \mathbb{R}^2$ defined by

$$G(f) = \left(\int_0^1 f(t)dt, \int_0^2 tf(t)dt \right) \quad \forall f \in V$$

It can be shown that G is a linear transformation and

$$G(1) = (1, 2) \quad , \quad G(t) = (1/2, 8/3) \quad , \quad G(t^2) = (1/3, 4)$$

so under the standard basis for the output space $\mathbb{R}^2$ we have

$$Mat(G) = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 2 & 8/3 & 4 \end{bmatrix}$$

6.14.1. Linear combinations of linear transformations

Let U and V be vector spaces over a common field $\mathbb{F}$. Let $T : U \to V$ and $G : U \to V$ be linear transformations. Fix scalars $a, b \in F$ and define $aT + bG$ as the function

$$(aT + bG)(x) = aT(x) + bG(x) \quad \forall x \in U$$

It can be shown that $aT + bG$ is a linear transformation. [To prove this, it may be easiest to define it by some letter, say, $H = aT + bG$, then show $H(\lambda x + \mu y) = \lambda H(x) + \mu H(y)$ for all $\lambda, \mu \in \mathbb{F}, x, y \in U$.] If $\dim(U) = n$ and $\dim(V) = m$ for some $n, m \in \mathbb{N}$, and if $\{u_1, ..., u_n\}$ and

$\{v_1, ..., v_m\}$ are bases for U and V, respectively, it holds that

$$Mat(aT + bG) = aMat(T) + bMat(G) \tag{6.38}$$

where the matrices are with respect to the given bases. An example is in Exercise 6.61.

Given U, V and their bases $\{u_1, ..., u_n\}$ and $\{v_1, ..., v_m\}$, one can define J as the set of all linear transformations $T : U \to V$. It can be shown that J is a vector space over field $\mathbb{F}$. [Linear combinations of two such linear transformations gives another linear transformation, and the 8 axioms hold.] Consider the map $Mat : J \to \mathbb{F}^{m \times n}$ that maps each $T \in J$ to its $m \times n$ matrix $Mat(T)$ with respect to the given bases. The equation (6.38) shows that Mat is itself a linear transformation.

6.14.2. Compositions of linear transformations

Let U, V, W be three vector spaces over a common field $\mathbb{F}$. Let $T : U \to V$ and $G : V \to W$ be linear transformations. Define the function $H : U \to W$ by

$$H(x) = G(T(x)) \quad \forall x \in U$$

H is the *composition* of functions G and T. It can be written $H = G \circ T$ or $H = GT$. It can be shown that H is a linear transformation. If $\dim(U) = n$, $\dim(V) = m$, $\dim(W) = k$ for some $n, m, k \in \mathbb{N}$, then under bases $\{u_1, ..., u_n\}$, $\{v_1, ..., v_m\}$, $\{w_1, ..., w_k\}$ for U, V, W, respectively, it can be shown that

$$Mat(G \circ T) = Mat(G)Mat(T)$$

An example is in Exercise 6.62.

6.15. Rank-nullity theorem for linear transformations

THEOREM 6.2. *(rank-nullity for linear transformation) Let U and V be vector spaces over a field $\mathbb{F}$. Suppose $T : U \to V$ is a linear transformation and $\dim(U) < \infty$. Then $rank(T) < \infty$ and*

$$\boxed{rank(T) + \dim(Null(T)) = \dim(U)}$$

In particular, define $r = rank(T)$, $m = \dim(Null(T))$, $n = \dim(U)$, and suppose

- *$\{y_1, ..., y_r\}$ is a basis for $Span(T)$.*
- *$\{x_1, ..., x_r\}$ are vectors in U such that $T(x_i) = y_i$ for all $i \in \{1, ..., r\}$.*
- *$\{z_1, ..., z_m\}$ is a basis for $Null(T)$.*

Then $\{x_1, ..., x_r, z_1, ..., z_m\}$ is a basis for U, so $r + m = n$.

PROOF. Exercise 6.5 shows that $\text{rank}(T) \leq \dim(U) < \infty$. If $r \geq 1$ and $m \geq 1$, it can be shown that $\{x_1, ..., x_r, z_1, ..., z_m\}$ are linearly independent in U (see Exercise 6.6) and span U (see Exercise 6.7). So these $r + m$ vectors are a basis for the n-dimensional space U, so $r + m = n$. The case $r = 0$ or $m = 0$ is left as an exercise. □

The above theorem assumes $\dim(U) < \infty$. Exercise 6.27 shows the rank-nullity equation still holds if $\dim(U) = \infty$:

$$rank(T) + \dim Null(T) = \infty$$

which means either $\text{rank}(T) = \infty$ or $\dim Null(T) = \infty$ (or both). This uses the standard "arithmetic with infinity":

$$\begin{aligned}
\infty + \infty = x + \infty = \infty + x = \infty \quad & \forall x \in \mathbb{R} \\
-\infty + -\infty = x + -\infty = -\infty + x = -\infty \quad & \forall x \in \mathbb{R} \\
\infty + -\infty &= \text{undefined} \\
-\infty + \infty &= \text{undefined}
\end{aligned}$$

[Caveat: Fix $x \in \mathbb{R}$. Within the rules of arithmetic with infinity, the equation $x + \infty = \infty$ is true, but we cannot "cancel ∞ from both sides" to conclude $x = 0$ because the undefined expression "$\infty + -\infty$" is not 0.]

6.16. Direct sums

Let V be a vector space over a field $\mathbb{F}$. Let $S_1, ..., S_k$ be subspaces. The sum $S_1 + S_2 + \cdots + S_k$ is defined according to the Minkowski set sum definition:

$$S_1 + S_2 + \cdots + S_k = \left\{ \sum_{i=1}^{k} s_i : s_i \in S_i \quad \forall i \in \{1, ..., k\} \right\}$$

Let S be another subspace of V. We say that S is a *direct sum* of $S_1, ..., S_k$, written

$$S = S_1 \oplus S_2 \oplus \cdots \oplus S_k \tag{6.39}$$

if $S = S_1 + S_2 + \cdots + S_k$, and every $s \in S$ can be *uniquely* written as

$$s = \sum_{i=1}^{k} s_i$$

for some unique $s_i \in S_i$ for $i \in \{1, \ldots, k\}$. Exercise 6.52 shows that S is a direct sum of $S_1, \ldots, S_k$ if and only if $S = \sum_{i=1}^{k} S_i$ and the following property holds:

$$\left(\sum_{i=1}^{k} v_i = 0, v_i \in S_i \quad \forall i \in \{1, \ldots, k\} \right) \Longrightarrow (v_i = 0 \quad \forall i \in \{1, \ldots, k\})$$

For example, suppose $\{v_1, v_2, v_3, v_4, v_5\}$ are linearly independent in V. Define

$$U = Span\{v_1, v_2\}$$
$$W = Span\{v_3, v_4, v_5\}$$

We claim:

$$Span\{v_1, v_2, v_3, v_4, v_5\} = U \oplus W$$

To prove this claim, first observe that $Span\{v_1, v_2, v_3, v_4, v_5\} = U + W$. Now suppose we have vectors $u \in U, w \in W$ such that

$$u + w = 0 \tag{6.40}$$

We want to show $u = 0$ and $w = 0$. Since $u \in U$ and $w \in W$ we have

$$u = \sum_{i=1}^{2} a_i v_i \quad \text{(for some scalars } a_1, a_2) \tag{6.41}$$

$$w = \sum_{i=3}^{5} a_i v_i \quad \text{(for some scalars } a_3, a_4, a_5) \tag{6.42}$$

From (6.40) we obtain

$$0 = u + w = \sum_{i=1}^{5} a_i v_i$$

which implies $a_i = 0$ for all $i \in \{1, \ldots, 5\}$ because $\{v_1, \ldots, v_5\}$ are linearly independent. Substituting $a_i = 0$ into (6.41)-(6.42) gives $u = 0$ and $w = 0$.

LEMMA 6.6. *Suppose S_1 and S_2 are subspaces of a vector space V. Then*

$$(S_1 + S_2 = S_1 \oplus S_2) \iff (S_1 \cap S_2 = \{0\})$$

PROOF. For the forward direction, suppose $S_1 + S_2 = S_1 \oplus S_2$. Then every vector in $S_1 + S_2$ can *uniquely* be written as the sum of a vector in S_1 and a vector in S_2. We now show that $S_1 \cap S_2 = \{0\}$. Suppose not (we reach a contradiction). Then there is a nonzero vector

$y \in S_1 \cap S_2$. Then

$$y = \underbrace{0}_{\in S_1} + \underbrace{y}_{\in S_2} = \underbrace{y}_{\in S_1} + \underbrace{0}_{\in S_2}$$

This gives two distinct ways of writing y as a sum of a vector in S_1 and a vector in S_2, a contradiction.

For the reverse direction, suppose $S_1 \cap S_2 = \{0\}$. We want to show that every vector in $S_1 + S_2$ can be *uniquely* written as the sum of a vector in S_1 and a vector in S_2. Fix $v \in S_1 + S_2$ and suppose $x_1, y_1 \in S_1$, $x_2, y_2 \in S_2$ and

$$v = x_1 + x_2 = y_1 + y_2$$

Thus, there are two ways of writing v. We want to show these two ways are in fact the same, so $x_1 = y_1$ and $x_2 = y_2$. Rearranging terms gives

$$\underbrace{x_1 - y_1}_{\in S_1} = \underbrace{y_2 - x_2}_{\in S_2}$$

which means that $x_1 - y_1 \in S_1 \cap S_2$. But the only vector in $S_1 \cap S_2$ is the vector 0, so $0 = x_1 - y_1$. Thus, $x_1 = y_1$. Since $0 = x_1 - y_1 = y_2 - x_2$, we also know $x_2 = y_2$. □

LEMMA 6.7. *Suppose $S, S_1, \ldots, S_k$ are subspaces of a vector space V and*

$$S = S_1 \oplus S_2 \oplus \cdots \oplus S_k$$

Then

a) $S_i \cap S_j = \{0\}$ for all $i \neq j$.

b) For each $i \in \{1, \ldots, k\}$ let $\{v_1[i], \ldots, v_{n[i]}[i]\}$ be linearly independent vectors in S_i (where $n[i] \in \mathbb{N}$ for all $i \in \{1, \ldots, k\}$). Then

$$\{v_1[1], \ldots, v_{n[1]}[1], v_1[2], \ldots, v_{n[2]}[2], \ldots, v_1[k], \ldots, v_{n[k]}[k]\}$$

are linearly independent vectors in S.

c) If $\dim(S_i) < \infty$ for all $i \in \{1, \ldots, k\}$ then $\dim(S) < \infty$ and

$$\dim(S) = \sum_{i=1}^{k} \dim(S_i)$$

PROOF. See Exercise 6.33. □

6.17. Dual space

Let V be a vector space over field $\mathbb{F}$. Define V^* as the set of all linear transformations $T : V \to \mathbb{F}$. It can be shown that V^* is itself a vector space, where vectors are linear transformations. Indeed, we can add linear transformations to produce new ones, we can scale linear transformations by scalars, and the 8 axioms for a vector space hold. The zero vector on the space V^* is the always-zero function that maps everything to the scalar 0. The vector space V^* is called the *dual space* for V.

What is $\dim(V^*)$? It turns out that if $\dim(V) < \infty$, then $\dim(V) = \dim(V^*)$.

- Case 1: Suppose $V = \{0\}$. Then the only linear transformation from $\{0\}$ to $\{0\}$ is the zero function. So $\dim(V) = \dim(V^*) = 0$.
- Case 2: Suppose $\dim(V) = n \in \mathbb{N}$. Let $\{v_1, ..., v_n\}$ be a basis for V. Then each $v \in V$ can be mapped to its unique vector of coefficients $a = (a_1, ..., a_n) \in \mathbb{F}^n$ such that

$$v = \sum_{i=1}^{n} a_i v_i \tag{6.43}$$

  For each $i \in \{1, ..., n\}$, define $T_i : V \to \mathbb{F}$ by mapping each $v \in V$ to the unique value a_i from (6.43). It can be shown that each T_i function is indeed a linear transformation (so it is a vector in V^*). Further, $\{T_1, ..., T_n\}$ is in fact a basis for V^*, since they can be shown to be linearly independent, and every linear transformation $T : V \to \mathbb{F}$ can be shown to be a linear combination of $\{T_1, ..., T_n\}$ (recall that a linear transformation is defined by its values on a basis for its domain). So $\dim(V) = \dim(V^*) = n$.

If $\dim(V) = \infty$, Appendix E shows that $\dim(V^*) = \infty$. [Here we are not distinguishing between the different levels of infinity.]

6.18. Dimension summary

We categorize a subspace S as either *finite dimensional* ($\dim(S) < \infty$) or *infinite dimensional* ($\dim(S) = \infty$). For the finite dimensional case, we define the *dimension* of S, written $\dim(S)$, as a nonnegative integer equal to the number of vectors in a basis for S:

$$(\dim(S) = 0) \iff (S = \{0\}) \quad \text{[the trivial subspace has no basis]}$$
$$(\dim(S) = n \in \mathbb{N}) \iff (S \text{ has a basis of size } n)$$
$$(\dim(S) = \infty) \iff (S \text{ is infinite dimensional})$$

$\dim(S) = \infty$ means S cannot be spanned by a finite number of vectors. We have not defined a basis for S when $\dim(S) = \infty$. One definition for this is a *Hamel basis* (Appendix E).

6.19. Exercises

EXERCISE 6.1. *(Is it a subspace?) Let V be the set of all real-valued functions $f(x)$ defined on domain $x \in (0, \infty)$.*

a) Let $S \subseteq V$ be the set of differentiable functions $f(x)$ that satisfy $f'(x) = 3f(x)^2$. Is S a subspace?

b) Let $S \subseteq V$ be the set of twice differentiable functions $f(x)$ that satisfy $f''(x) = 12f(x) - f'(x)$. Is S a subspace?

EXERCISE 6.2. *(Are they linearly independent?) Suppose x, y, z are vectors that satisfy*

- *$\{x, y\}$ are linearly independent.*
- *$\{y, z\}$ are linearly independent.*
- *$\{x, z\}$ are linearly independent.*

True or false: $\{x, y, z\}$ are linearly independent. Prove true, or provide a counter-example.

EXERCISE 6.3. *Fix $\mathbb{F} = \mathbb{R}$ and let V be the vector space of all functions $f : \mathbb{R} \to \mathbb{R}$. Let $S \subseteq V$ be the set of all functions that satisfy $\int_{-\infty}^{\infty} f(t)^2 dt < \infty$, called* square integrable *functions.*[3] *Show that S is a subspace of V. Hence, S can itself be viewed as a vector space (over the field $\mathbb{R}$). Hint: For any functions $f, g \in V$ we have $(f(t) + g(t))^2 \leq 2f(t)^2 + 2g(t)^2$ for all $t \in \mathbb{R}$.*

EXERCISE 6.4. *Suppose U, V are vector spaces over a field $\mathbb{F}$. Suppose $T : U \to V$ is a linear transformation.*

a) Show that $Null(T)$ is a subspace of U.

b) Show that $Span(T)$ is a subspace of V.

EXERCISE 6.5. *Suppose U, V are vector spaces over a field $\mathbb{F}$. Let $T : U \to V$ be a linear transformation. Suppose $\dim(U) = n$ for some positive integer n and let $\{x_1, ..., x_n\}$ be a basis for U. The vector space V is allowed to be infinite dimensional.*

a) Show $Span\{T(x_1), T(x_2), ..., T(x_n)\} = V$, so $\dim(Span(T)) \leq n$.

b) Explain why $\{T(x_1), T(x_2), ..., T(x_n)\}$ are not necessarily a basis for $Span(T)$.

EXERCISE 6.6. *Suppose U and V are vector spaces over some field $\mathbb{F}$. Suppose $\dim(U) = n$ for some positive integer n. Let $T : U \to V$ be a linear transformation. Fix positive integers r, m and suppose*

- *$\{y_1, ..., y_r\}$ is a basis for $Span(T)$.*
- *$\{x_1, ..., x_r\}$ are vectors in U such that $T(x_i) = y_i$ for all $i \in \{1, ..., r\}$.*
- *$\{z_1, ..., z_m\}$ is a basis for $Null(T)$.*

[3] This problem treats square integrable functions over the domain $\mathbb{R}$. Functions with domain $[a, b]$ are said to be *square integrable* if $\int_a^b f(t)^2 dt < \infty$. Measure theory detail: To be precise, we use the Lebesgue integral. For the integral to make sense, we implicitly require a square integrable function to be Borel measurable.

Show that $\{x_1, ..., x_r, z_1, ..., z_m\}$ are linearly independent vectors in U. Hint: Start by writing the first sentence of the proof.

EXERCISE 6.7. *For the same assumptions as Exercise 6.6, we want to show $\{x_1, ..., x_r, z_1, ..., z_m\}$ span U. Fix $u \in U$.*

a) Explain why $T(u) = \sum_{i=1}^{r} c_i y_i$ for some scalars c_i.

b) Define $x = \sum_{i=1}^{r} c_i x_i$. Explain why $x \in U$, $u \in U$, and $T(x) = T(u)$.

c) Show that $u - x \in Null(T)$. Conclude that u is a linear combination of $\{x_1, ..., x_r, z_1, ..., z_m\}$.

EXERCISE 6.8. *(Linear independence in the space of functions) Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$.*

a) Prove that $\{\cos(t), \sin(t), t\}$ are linearly independent.

b) Prove that $\{t, t + 2, e^t\}$ are linearly independent.

EXERCISE 6.9. *(More linearly independent functions) Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$. Prove the following collections are linearly independent, or provide a specific counter-example to show they are linearly dependent.*

a) $\{4t^2, 5t - 3, 3t + 4, 1 + t^2\}$

b) $\left\{t, te^t, \frac{t^3+1}{t^2+2}e^t\right\}$

c) $\{\sin(t)^2, \frac{1}{2}\cos(t)^2, 5\}$

EXERCISE 6.10. *Let V be the vector space of all functions $f : \mathbb{R} \to \mathbb{R}$. Define the linear operator $T : V \to V$ by $T(f)(t) = \sin(t) f(t)$ for all $t \in \mathbb{R}$. That is, T takes a real-valued function f and forms a new function by multiplying $f(t)$ by $\sin(t)$ for all $t \in \mathbb{R}$.*

a) Show $Null(T)$ is the set of all $f \in V$ such that $f(t) = 0$ whenever $t \notin \{\pi k\}_{k=-\infty}^{\infty}$.

b) Show $Span(T)$ is the set of all $g \in V$ such that $g(t) = 0$ whenever $t \in \{\pi k\}_{k=-\infty}^{\infty}$

EXERCISE 6.11. *(Sampling a nonlinear function) Consider the following nonlinear function T from $\mathbb{R}^3$ to $\mathbb{R}^3$:*

$$T(\vec{x}) = A\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} + \begin{bmatrix} x_1^2 + 2 \\ x_2 \\ x_2 x_3 \end{bmatrix} \quad \forall \vec{x} \in \mathbb{R}^3$$

where A is an unknown 3×3 matrix and $\vec{x} = (x_1, x_2, x_3)$. Find A from the following information: We sample this function at vectors $\vec{x}_1, \vec{x}_2, \vec{x}_3$ and find $T(\vec{x}_1) = \vec{v}_1$, $T(\vec{x}_2) = \vec{v}_2$, $T(\vec{x}_3) = \vec{v}_3$, where:

$$\vec{x}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, \vec{x}_2 = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \vec{x}_3 = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

$$\vec{v}_1 = \begin{bmatrix} -1 \\ 0 \\ 2 \end{bmatrix}, \vec{v}_2 = \begin{bmatrix} 2 \\ 2 \\ 2 \end{bmatrix}, \vec{v}_3 = \begin{bmatrix} 1 \\ 1 \\ -2 \end{bmatrix}$$

EXERCISE 6.12. *Design a problem similar to the previous one.*

EXERCISE 6.13. *(Subsets) Let V be a vector space over a field $\mathbb{F}$. Let S_1, S_2 be subspaces of V that satisfy $S_1 \subseteq S_2$ and $\dim(S_1) = \dim(S_2) < \infty$. Show that $S_1 = S_2$*

EXERCISE 6.14. *(If $rank(T) = rank(T^2)$) Let V be a finite-dimensional vector space over a field. Suppose $T : V \to V$ is a linear transformation that satisfies $rank(T) = rank(T^2)$. Prove that $Null(T) = Null(T^2)$.*

EXERCISE 6.15. *(If $Null(T^2) = Null(T)$) Suppose $T : V \to V$ is a linear transformation that satisfies $Null(T^2) = Null(T)$. Prove that $Null(T^k) = Null(T)$ for all $k \in \{1, 2, 3, ...\}$.*

EXERCISE 6.16. *(Finite and infinite dimensional spaces)*

a) Let V be a finite-dimensional vector space over a field $\mathbb{F}$. Let $T : V \to V$ be a linear transformation such that $Null(T) = \{0\}$. Show that $Span(T) = V$.

b) Give an example of a vector space V and a linear transformation $T : V \to V$ such that $Null(T) = \{0\}$, but $Span(T)$ is not the full vector space V.

EXERCISE 6.17. *(Spanning $T(S)$) Let U and V be vector spaces over a field $\mathbb{F}$. Suppose*

- *$T : U \to V$ is a linear transformation.*
- *S is a subspace of U.*
- *$\{s_1, s_2, ..., s_k\}$ are vectors in U that span S.*

a) Define the subspace $T(S) = \{T(s) \in V : s \in S\}$. Show that $\{T(s_1), T(s_2), ..., T(s_k)\}$ are vectors in V that span $T(S)$.

b) Conclude that if $\dim(U) < \infty$ then $\dim(T(U)) < \infty$.

EXERCISE 6.18. *(Linear independence of polynomials) Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$.*

a) Use induction to prove that for all positive integers k, the polynomial functions $\{1, t, t^2, ..., t^k\}$ are linearly independent.

b) What is the dimension of the subspace of all polynomials of degree less than or equal to 10?

EXERCISE 6.19. *(Linear dependence of functions) Consider three functions $f_i : \mathbb{R} \to \mathbb{R}$ defined as*

$$f_1(t) = 7t - 1 \quad \forall t \in \mathbb{R}$$
$$f_2(t) = 2t + 8 \quad \forall t \in \mathbb{R}$$

$$f_3(t) = 10t - 5 \quad \forall t \in \mathbb{R}$$

Show $\{f_1, f_2, f_3\}$ are linearly dependent by explicitly writing one of the functions as a linear combination of the other two.

EXERCISE 6.20. *(First sentence of linear independence proof) Write the first 2 sentences of the proof. Use good mathematical words such as "fix," "suppose," "for some," "for all." Make sure your second sentence has the form "We want to show..."*

a) Prove $\{x_1, ..., x_k\}$ are linearly independent.

b) Prove $\{x, y, z\}$ are linearly independent.

c) Prove $\{e^t, t, 1, h(t)\}$ are linearly independent.

EXERCISE 6.21. *(Proving that something satisfies the requirements of a definition) Write the first 2 sentences of the proof. Use good mathematical words such as "fix," "suppose," "for some," "for all." Make sure your second sentence has the form "We want to show..."*

a) Prove $T : U \to V$ is a linear transformation.

b) Prove S is a subspace.

c) Prove $S \subseteq Col(A)$.

d) Prove $S \subseteq Span\{v_1, ..., v_k\}$.

e) Prove $S \subseteq H^\perp$.

f) Prove $S \subseteq Col(A)^\perp$.

EXERCISE 6.22. *(True/False) All subspaces considered below are subsets of some common vector space V. State whether true or false. Explain why.*

a) The union of subspaces is a subspace.

b) The intersection of subspaces is a subspace.

c) The sum of two subspaces (also called a Minkowski set sum) is again a subspace.

d) Adding three vectors in a subspace gives another vector in that subspace.

e) Adding a constant to a vector in a subspace gives another vector in that subspace.

f) If $\{x_1, x_2, x_3, ...\}$ is an infinite sequence of vectors in a set $D \subseteq V$, then $\sum_{i=1}^{\infty}(1/2^i)x_i$ is a linear combination of vectors in D.

EXERCISE 6.23. *(Pairwise relationships) Suppose x, y, z are vectors in some vector space V. Prove the following claim, or provide a counter-example. Claim: If $\{x, y\}$ are linearly independent, and $\{y, z\}$ are linearly independent, and $\{x, z\}$ are linearly independent, then $\{x, y, z\}$ are linearly independent.*

EXERCISE 6.24. *(Shifted sets) Let V be a vector space. Let S be a subset of V. Fix $w \in V$. Define the* shifted set

$$w + S = \{w + s : s \in S\}$$

In the special case when S is a subspace, the shifted set $w + S$ is called an affine set.

a) Show that if S is a subspace and $w \in S$, then $w + S = S$.

b) Give an example when the subspace $S \subseteq \mathbb{R}^2$ and $w \in S$. Draw the picture of S and $w + S$.

c) For your same subspace of part (b), now give an example with $w \notin S$. Draw the picture of S and $w + S$.

EXERCISE 6.25. *(Affine sets) Let V be a vector space over a field $\mathbb{F}$. A set $X \subseteq V$ is said to be* affine *if $X = w + S$ for some $w \in V$ and some subspace $S \subseteq V$.*

a) Prove that a set $X \subseteq V$ is affine if and only if it is nonempty and $\sum_{i=1}^{n} a_i x_i \in X$ for all positive integers n, all $x_1, \ldots, x_n \in X$, and all scalars $a_1, \ldots, a_n$ such that $\sum_{i=1}^{n} a_i = 1$.

b) Show that X is affine if and only if $ax + (1 - a)y \in X$ for all $x, y \in X$ and all scalars a.

EXERCISE 6.26. *(Multiple choice) Let S be a subspace. What are the possible values of $\dim(S)$?*

a) $\{1, 2, 3, \ldots\}$

b) $\{0, 1, 2, 3, \ldots\}$

c) $\{\infty\} \cup \{0, 1, 2, 3, \ldots\}$

d) $\{0, 1, 2, 3, \ldots\} \cup \{-1, -2, -3, \ldots\}$

e) $\{-\infty\} \cup \{5, 42\}$

EXERCISE 6.27. *(Rank-nullity for infinite-dimensional domain) Let U and V be vector spaces over a field $\mathbb{F}$. Let $T : U \to V$ be a linear transformation. Suppose $\dim(U) = \infty$. Section 6.6 ensures there is an infinite sequence $\{x_1, x_2, x_3, \ldots\}$ of vectors in U such that every finite subcollection of these vectors is linearly independent. For $n \in \mathbb{N}$ define*

$$U_n = Span\{x_1, \ldots, x_n\}$$

and define $T_n : U_n \to V$ by $T_n(x) = T(x)$ for all $x \in U_n$.

a) Show that T_n is a linear transformation on a finite dimensional domain and so (by Theorem 6.2):

$$rank(T_n) + \dim Null(T_n) = n \quad \forall n \in \mathbb{N}$$

b) Show that for all $n \in \mathbb{N}$ we have

$$Null(T_n) \subseteq Null(T)$$
$$rank(T_n) \leq rank(T)$$

c) Show that either $rank(T) = \infty$ or $\dim Null(T) = \infty$ (and so $rank(T) + \dim Null(T) = \infty$).

EXERCISE 6.28. *(Example linear transformations with infinite domain)*

a) Give a specific example of a linear transformation $T : U \to V$ with $rank(T) < \infty$ and $\dim Null(T) = \infty$.

b) Give a specific example of a linear transformation $T : U \to V$ with $rank(T) = \infty$ and $\dim Null(T) < \infty$.

c) Give a specific example of a linear transformation $T : U \to V$ with $rank(T) = \infty$ and $\dim Null(T) = \infty$.

EXERCISE 6.29. *(A subspace S that is not closed) Let V be the vector space of continuous functions $f : [0,1] \to \mathbb{R}$. Let C be the infinite sequence of functions*

$$C = \{1, t, t^2, t^3, ...\}$$

Let S be the subspace of all linear combinations of functions in C.

a) Show that, because a linear combination is required to be a finite sum*, the set S consists precisely of all polynomial functions $f(t) = a_0 + a_1 t + \cdots + a_k t^k$ for some integer $k \geq 0$ and some real numbers $a_0, ..., a_k$.*

b) For each $n \in \mathbb{N}$, define $g_n : [0,1] \to \mathbb{R}$ by

$$g_n(t) = \sum_{i=1}^{n} (1/2)^i t^i \quad \forall t \in [0,1]$$

Then $g_n \in S$ for each $n \in \mathbb{N}$. Use the geometric sum formula to get a closed form expression for g_n.

c) Consider the function $g : [0,1] \to \mathbb{R}$ defined by

$$g(t) = \sum_{i=1}^{\infty} (1/2)^i t^i \quad \forall t \in [0,1]$$

For each $t \in [0,1]$, it can be shown that $\sum_{i=1}^{n} (1/2)^i t^i$ converges to a real number as $n \to \infty$, so the above function g is well defined. In particular

$$g(t) = \lim_{n \to \infty} g_n(t) \quad \forall t \in [0,1]$$

The function g is defined by an infinite sum, while linear combinations of functions in C are required to be finite sums, so g is not necessarily in S. It is only in S if the infinite sum happens to converge to a polynomial. Take the limit of functions in part (b) as $n \to \infty$ to obtain a closed form expression for g. Show that g is not a polynomial, hence, $g \notin S$. [V is a real inner product space *(defined in Chapter 8). Using the standard* inner product norm*, it can be shown $||g_n - g|| \to 0$, so g is a limit point of S but is not in S, so S is not a closed set. Subspace S is infinite dimensional. It can be shown that every finite dimensional subspace of V is closed (Exercise 8.30).]*

EXERCISE 6.30. *($\mathbb{F}^{\mathbb{N}}$ cannot be spanned by a mere countably infinite collection of vectors) Let $\mathbb{F}$ be a field. Recall that $\mathbb{F}^{\mathbb{N}}$ denotes the vector space of all infinite sequences of scalars in $\mathbb{F}$:*

$$\mathbb{F}^{\mathbb{N}} = \{(x_1, x_2, x_3, ...) : x_i \in \mathbb{F} \quad \forall i \in \mathbb{N}\}$$

It is clear that $\mathbb{F}^{\mathbb{N}}$ is infinite dimensional, so it cannot be spanned by any finite list of vectors.

This problem shows that $\mathbb{F}^{\mathbb{N}}$ cannot even be spanned by any countably infinite list of vectors. Central to this result is the fact that a linear combination of vectors in a collection C is always a finite sum.

a) Define a countably infinite collection $E = \{e_1, e_2, e_3, ...\}$ of vectors in $\mathbb{F}^{\mathbb{N}}$ according to the following naive guess *of a spanning set:*

$$e_1 = (1, 0, 0, 0, ...)$$
$$e_2 = (0, 1, 0, 0, ...)$$
$$e_3 = (0, 0, 1, 0, ...)$$

and so on. So for each $i \in \mathbb{N}$, e_i is a sequence that is all-zero except in entry i (where it is 1). Define S as the subspace of $\mathbb{F}^{\mathbb{N}}$ that consists of all linear combinations of vectors in E. Explain why $S \neq \mathbb{F}^{\mathbb{N}}$. To do this, you should find a vector in $\mathbb{F}^{\mathbb{N}}$ that is not in S.

b) Suppose $C = \{b_1, b_2, b_3, ...\}$ is some countably infinite list of vectors in $\mathbb{F}^{\mathbb{N}}$. Define S as the subspace of all linear combinations of vectors in C. Then

$$S = \cup_{k=1}^{\infty} Span\{b_1, ..., b_k\}$$

We manufacture a vector $w \in \mathbb{F}^{\mathbb{N}}$ such that $w \notin S$. We first make a convenient definition: For every $v = (v_1, v_2, v_3, ...) \in \mathbb{F}^{\mathbb{N}}$ and all positive integers i, j with $i < j$, define $v[i, j]$ as

$$v[i, j] = (v_i, ..., v_j) \in \mathbb{F}^{j-i+1}$$

For example

$$v = (v_1, v_2, v_3, v_4, v_5, ...) \implies v[2, 4] = (v_2, v_3, v_4) \in \mathbb{F}^3$$

Observe that for all $i, j, k \in \mathbb{N}$ with $i < j$ we have

$$(v \in Span\{b_1, ..., b_k\}) \implies v[i, j] \in Span\{b_1[i, j], ..., b_k[i, j]\}$$

Now construct $w = (w_1, w_2, w_3, ...) \in \mathbb{F}^{\mathbb{N}}$ as follows:

- *Step 1: Choose $(w_1, w_2) \in \mathbb{F}^2$ as any vector that is not in $Span\{b_1[1, 2]\}$. That is, we choose a vector in $\mathbb{F}^2$ that is not a scalar multiple of a given vector in $\mathbb{F}^2$.*
- *Step 2: Choose $(w_3, w_4, w_5) \in \mathbb{F}^3$ as any vector that is not in $Span\{b_1[3, 5], b_2[3, 5]\}$. That is, we choose a vector in $\mathbb{F}^3$ that is not in the span of two given vectors in $\mathbb{F}^3$.*
- *Step 3: Choose $(w_6, w_7, w_8, w_9) \in \mathbb{F}^4$ as any vector not in $Span\{b_1[6, 9], b_2[6, 9], b_3[6, 9]\}$.*

And so on. Give the general step k for $k \in \mathbb{N}$. Explain why at each step k, the desired vector in $\mathbb{F}^{k+1}$ exists. Then explain why the constructed vector $w \in \mathbb{F}^{\mathbb{N}}$ satisfies $w \notin S$.

EXERCISE 6.31. *(Hamel basis) Let S be a nonempty and possibly infinite dimensional subspace of some vector space V over a field $\mathbb{F}$. A* Hamel basis *for S is a collection C of vectors in S such that: (i) Every finite subcollection of vectors in C forms a linearly independent set; (ii) Every vector in S can be written as a linear combination of vectors in C. Central to this definition is the fact that linear combinations of vectors in C are always* finite sums. *The* axiom of choice *can be used to show that every nontrivial subspace has a Hamel basis.*

a) Let V be the vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$. Define the following countably infinite list of vectors in V:

$$C = \{1, t, t^2, t^3, \ldots\}$$

Show that C is a Hamel basis for V.

b) Let $\mathbb{F}$ be a field. Let $V = \mathbb{F}^{\mathbb{N}}$ be the vector space of all countably infinite sequences of scalars in $\mathbb{F}$. Explain why the result of Exercise 6.30 implies that any Hamel basis for $\mathbb{F}^{\mathbb{N}}$ must be uncountably infinite.

c) Define V as the real vector space of continuous functions $f : \mathbb{R} \to \mathbb{R}$. The result of Exercise 6.30 implies that $\mathbb{R}^{\mathbb{N}}$ cannot be spanned by a countable list of vectors. Use this to show that any Hamel basis of V is uncountably infinite.

EXERCISE 6.32. *(Proof from vector space axioms) Prove Lemma 6.1.*

EXERCISE 6.33. *Prove Lemma 6.7. Hint: For part (a), reuse the proof technique in Lemma 6.6. For part (c), consider a basis for each S_i and use part (b).*

EXERCISE 6.34. *(A vector space of matrices) Let $\mathbb{F}$ be a field. Let V be the vector space of 2×2 matrices with entries in $\mathbb{F}$.*

a) Let $S_1 \subseteq V$ be the set of all 2×2 invertible matrices. Is S_1 a subspace of V?

b) Let $S_2 \subseteq V$ be the set of all 2×2 upper triangular matrices. Is S_2 a subspace of V?

c) Let $S_3 \subseteq V$ be the set of all 2×2 matrices A with $\text{rank}(A) \leq 1$. Is S_3 a subspace of V?

EXERCISE 6.35. *(A subspace of $\mathbb{F}^{m \times n}$) Let $\mathbb{F}$ be a field. Let $\mathbb{F}^{m \times n}$ be the vector space of all $m \times n$ matrices with entries in $\mathbb{F}$. Fix $m, n \in \mathbb{N}$. Fix $B \in \mathbb{F}^{n \times n}$ and define*

$$S = \{A \in \mathbb{F}^{m \times n} : AB = 0\}$$

where 0 is the all-zero matrix of size $m \times n$.

a) Show that S is a subspace of $\mathbb{F}^{m \times n}$.

b) In terms of m and n, what is the dimension of $\mathbb{F}^{m \times n}$?

c) In terms of m and n, find the largest-possible and smallest-possible dimension of S. [Consider different possibilities for the square matrix B.]

EXERCISE 6.36. *(Random variables with finite second moments) Let $(\Omega, \mathcal{F}, P)$ be a probability*

space. Let V be the real vector space of random variables $X : \Omega \to \mathbb{R}$. Define

$$S_1 = \{X \in V : X(\omega) \geq 0 \quad \forall \omega \in \Omega\}$$
$$S_2 = \left\{X \in V : \mathbb{E}\left[X^2\right] < \infty\right\}$$

a) Prove that S_1 is a subspace, or find a counter-example.

b) Prove that S_2 is a subspace, or find a counter-example.

EXERCISE 6.37. *(Two flips of a coin) Consider a probability experiment consisting of two flips of a fair coin. Let $\Omega = \{(H,H),(H,T),(T,H),(T,T)\}$ be the sample space. Let $\mathcal{F}$ be the set of all subsets of Ω. Let $P : \mathcal{F} \to [0,1]$ be the probability measure that makes all four outcomes equally likely, so $P[A] = |A|/4$ for all $A \subseteq \Omega$, where $|A|$ denotes the number of outcomes in set A. Define V as the set of all random variables on this probability space: $V = \{\text{functions } X : \Omega \to \mathbb{R}\}$. Define*

$$S_1 = \{X \in V : X(H,T) = X(T,H)\}$$
$$S_2 = \{X \in V : X(T,T) \geq 0\}$$
$$S_3 = \{X \in V : P[X \leq 7.2] \geq 1/2\}$$
$$S_4 = \{X \in V : \mathbb{E}[X] = 0\}$$

a) Is S_1 a subspace?

b) Is S_2 a subspace?

c) Is S_3 a subspace?

d) Give two examples of random variables in S_4. Is S_4 a subspace?

EXERCISE 6.38. *(Expectations and equally likely outcomes) Define $\Omega = \{red, blue, green\}$. Define $\mathcal{F}$ as the set of all subsets of Ω. Let $P : \mathcal{F} \to [0,1]$ be the probability measure that makes all three outcomes equally likely, so $P[A] = |A|/3$ for all $A \subseteq \Omega$, where $|A|$ is the size of set A. Define V as the real vector space of all random variables $X : \Omega \to \mathbb{R}$. Define $Y, Z \in V$ by*

$$Y(red) = 0, \quad Y(blue) = 1, \quad Y(green) = -1$$
$$Z(red) = 0, \quad Z(blue) = 2, \quad Z(green) = 4$$

Define the following subsets of V:

$$S_1 = \{X \in V : \mathbb{E}[X] = 0\}$$
$$S_2 = \{X \in V : X(red) = 0\}$$
$$S_3 = \left\{X \in V : 10\mathbb{E}[X] = 3\mathbb{E}\left[X^2\right]\right\}$$

a) Is $Y \in S_1$? Is $Z \in S_1$? Is S_1 a subspace?

b) Is $Y \in S_2$? Is $Z \in S_2$? Is S_2 a subspace?

c) Is $Y \in S_3$? Is $Z \in S_3$? Is S_3 a subspace?

d) Give an example random variable $W \in V$ such that $W \notin S_1$ and $W \notin S_2$, but $W \in S_1 + S_2$. [Here, the sum of sets uses the Minkowski set sum definition: $S_1 + S_2 = \{s_1 + s_2 : s_1 \in S_1, s_2 \in S_2\}$.]

EXERCISE 6.39. *Let V be a vector space over a field $\mathbb{F}$. Let S be a subspace. Find $S + S$. [Here, the sum of sets uses the Minkowski set sum definition: $S + S = \{s_1 + s_2 : s_1 \in S, s_2 \in S\}$.]*

EXERCISE 6.40. *(Direct sum for $\mathbb{R}^3$) Let S be a 2-dimensional subspace of $\mathbb{R}^3$. Fix $y \in \mathbb{R}^3$ such that $y \notin S$.*

a) Show that $\mathbb{R}^3 = S \oplus Span\{y\}$.

b) Give an example of this situation.

EXERCISE 6.41. *(Symmetry and skew symmetry) Let $\mathbb{F}$ be a field. Define $2 = 1 + 1$. [Examples of fields for which $2 \neq 0$ and $2 = 0$ are given in Appendix C.2.] Fix $n \in \mathbb{N}$. Let V be the vector space of all $n \times n$ matrices with entries in $\mathbb{F}$. Define*

$$S_1 = \{B \in V : B^\top = B\} \quad \textit{(symmetric)}$$
$$S_2 = \{B \in V : B^\top = -B\} \quad \textit{(skew symmetric)}$$

a) Show S_1 and S_2 are subspaces.

b) For every $A \in V$, show that $A + A^\top \in S_1$.

c) For every $A \in V$, show that $A - A^\top \in S_2$.

d) Suppose $\mathbb{F}$ is such that $2 \neq 0$. Find $S_1 + S_2$.

e) Suppose $\mathbb{F}$ is such that $2 = 0$. Find $S_1 + S_2$.

f) For $\mathbb{F} = \mathbb{R}$ and $n = 3$, give an example of a nonzero matrix in S_1, and another nonzero matrix in S_2.

EXERCISE 6.42. *Prove Lemma 6.3 using the independent-set-spanning-set comparison theorem (Theorem 6.1).*

EXERCISE 6.43. *Prove Lemma 6.4 using the independent-set-spanning-set comparison theorem (Theorem 6.1).*

EXERCISE 6.44. *Give an example of a vector space V over the field $\mathbb{R}$ and a subspace $S \subseteq V$ such that $\dim(S) = \dim(V)$, but $S \neq V$. Explain why this does not contradict Lemma 6.4.*

EXERCISE 6.45. *(Basis for symmetric 2×2 matrices) Let V be the real vector space of 2×2 matrices with entries in $\mathbb{R}$. Let S be the subspace of V consisting of 2×2 symmetric matrices.*

a) Show that $\dim(V) = 4$ by finding a basis of size 4.

b) Find a basis for S. What is $\dim(S)$?

EXERCISE 6.46. *(Basis for 3×3 and $n \times n$ symmetric matrices) Let $\mathbb{R}^{3\times 3}$ be the real vector space of 3×3 matrices with real entries. It holds that $\dim(\mathbb{R}^{3\times 3}) = 9$.*

a) Find a basis for the subspace of all 3×3 symmetric matrices. What is the dimension?

b) Suppose we consider the vector space $\mathbb{R}^{n\times n}$ of $n \times n$ matrices. Make a guess about the dimension of the subspace of $n \times n$ symmetric matrices. Explain your reasoning.

EXERCISE 6.47. *Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$.*

a) Show that $\{1, t(t-1), t(t-1)^2\}$ are linearly independent in V.

b) Show the vectors in (a) are a basis for $\{f \in V : f(0) = f(1)\}$.

EXERCISE 6.48. *(linear independence by discontinuity) Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$. Let $f \in V$ be a discontinuous function. Let $g \in V$ be a continuous function that is not the always-zero function. Show that $\{f, g\}$ are linearly independent.*

EXERCISE 6.49. *(Linear difference equation) Let $\mathbb{R}^{\mathbb{N}}$ be the set of infinite real-valued sequences $(x_1, x_2, x_3, ...)$. Define $S \subseteq \mathbb{R}^{\mathbb{N}}$ by*

$$S = \{(x_1, x_2, x_3, ...) \in \mathbb{R}^{\mathbb{N}} : x_{k+2} + x_{k+1} - 6x_k = 0 \quad \forall k \in \mathbb{N}\}$$

a) Show that S is a subspace.

b) Show that there is a unique sequence $(u_k) \in S$ such that $(u_1, u_2) = (1, 0)$. Show there is a unique sequence $(v_k) \in S$ such that $(v_1, v_2) = (0, 1)$. Argue that $S = Span\{(u_k), (v_k)\}$.

c) Show that $\{(u_k), (v_k)\}$ are linearly independent. Conclude $\dim(S) = 2$.

d) Find two linearly independent sequences in S by searching for solutions to the linear difference equation of the form $x_k = \theta^k$ for $k \in \mathbb{N}$, for some scalar θ.

e) Use part (d) to find an explicit form for (u_k) and (v_k).

EXERCISE 6.50. *(Linear difference equations with $\mathbb{F}$) Let $\mathbb{F}$ be a field. Consider the vector space $\mathbb{F}^{\mathbb{N}}$ of infinite sequences $(x_1, x_2, x_3, ...)$ with $x_i \in \mathbb{F}$ for all $i \in \mathbb{N}$. Fix $m \in \mathbb{N}$. Define $S \subseteq \mathbb{F}^{\mathbb{N}}$ by*

$$S = \left\{(x_1, x_2, x_3, ...) \in \mathbb{F}^{\mathbb{N}} : x_{k+m} + \sum_{i=0}^{m-1} a_i x_{k+i} = 0 \quad \forall k \in \mathbb{N}\right\}$$

where $a_0, ..., a_{m-1}$ are given scalars in $\mathbb{F}$.

a) Show S is a subspace.

b) Using arguments similar to those in Section 6.8.6, argue that for each $(\lambda_1, ..., \lambda_m) \in \mathbb{F}^m$, there is a unique sequence $(x_k) \in S$ such that $(x_1, ..., x_m) = (\lambda_1, ..., \lambda_m)$. In particular, the only sequence in S that has m leading zeros is the always-zero sequence.

c) Let r be a positive integer. Show that sequences $\{(x_k^{(1)}), ..., (x_k^{(r)})\}$ are linearly independent in S if and only if $\{(x_1^{(1)}, ..., x_m^{(1)}), ..., (x_1^{(r)}, ..., x_m^{(r)})\}$ are linearly independent in $\mathbb{F}^m$.

d) Conclude that $\dim(S) = m$.

EXERCISE 6.51. *(Left-shift by one) Let $\mathbb{F}$ be a field and let $\mathbb{F}^{\mathbb{N}}$ be the set of all infinite sequences $(x_1, x_2, x_3, ...)$ of scalars in $\mathbb{F}$. Define the linear operator $T : \mathbb{F}^{\mathbb{N}} \to \mathbb{F}^{\mathbb{N}}$ by*

$$T((x_1, x_2, x_3, ...)) = (x_2, x_3, x_4, ...) \quad \forall x \in \mathbb{F}^{\mathbb{N}}$$

a) Prove that $Span(T) = \mathbb{F}^{\mathbb{N}}$.

b) Find $Null(T)$. *What is* $\dim Null(T)$*?*

EXERCISE 6.52. *(Equivalent condition for direct sum) Let* V *be a vector space over a field* $\mathbb{F}$. *Let* $S, S_1, ..., S_k$ *be subspaces. Prove that* $S = S_1 \oplus S_2 \oplus \cdots \oplus S_k$ *if and only if* $S = \sum_{i=1}^{k} S_i$ *and the following property holds*

$$\left(\sum_{i=1}^{k} v_i = 0, v_i \in S_i \quad \forall i \in \{1, ..., k\} \right) \implies (v_i = 0 \quad \forall i \in \{1, ..., k\})$$

For the reverse direction, assume the uniqueness property fails, so there is a vector $x \in S$ *such that* $x = \sum_{i=1}^{k} s_i = \sum_{i=1}^{k} w_i$ *for* $s_i, w_i \in S_i$ *for* $i \in \{1, ..., k\}$*, but there is at least one* i *for which* $s_i \neq w_i$ *(then reach a contradiction).*

EXERCISE 6.53. *(A function satisfying the second property of linear transformation, but not the first) Consider the real vector space* $\mathbb{R}^2$. *Construct an example function* $T : \mathbb{R}^2 \to \mathbb{R}^2$ *that satisfies* $T(ax) = aT(x)$ *for all* $x \in \mathbb{R}^2$, $a \in \mathbb{R}$, *but is not a linear transformation because the property* $T(u + w) = T(u) + T(w)$ *is violated for some particular* $u, w \in \mathbb{R}^2$.

EXERCISE 6.54. *(A function satisfying the first property of linear transformation, but not the second) Consider the complex field* $\mathbb{C}$. *Treat* $\mathbb{C}$ *itself as a complex vector space. Consider the function* $T : \mathbb{C} \to \mathbb{C}$ *defined by*

$$T(a + ib) = a + 2ib \quad \forall a, b \in \mathbb{R}$$

a) Show that $T(x + y) = T(x) + T(y)$ *for all* $x, y \in \mathbb{C}$.

b) Show that $T(\theta u) \neq \theta T(u)$ *for some particular* $\theta \in \mathbb{C}$, $u \in \mathbb{C}$.

EXERCISE 6.55. *(Linear transformation with field* $\mathbb{Q}$*) Let* U, V *be vector spaces over the field* $\mathbb{Q}$ *of rational numbers. Let* $T : U \to V$ *be a function that satisfies the first property of a linear transformation:* $T(x + y) = T(x) + T(y)$ *for all* $x, y \in U$. *For this special case with field* $\mathbb{Q}$, *we show this is enough to ensure* T *is a linear transformation.*

a) Show that $T(0) = 0$.

b) Show that $T(nx) = nT(x)$ *for all* $x \in U$ *and all* $n \in \mathbb{N}$.

c) Show that $T(nx) = nT(x)$ *for all* $x \in U$ *and all integers* n *(positive, negative, or zero).*

d) Use part (c) to show that $T(\frac{p}{q}x) = \frac{p}{q}T(x)$ *for all* $x \in U$ *and all rational numbers* $p/q \in \mathbb{Q}$ *(so* p, q *are integers and* $q \neq 0$*).*

e) Explain how this idea can be used to show that if U, V *are vector spaces over the field* $\mathbb{R}$, *and if* $T : U \to V$ *is a function that satisfies* $T(x + y) = T(x) + T(y)$ *for all* $x, y \in U$, *then* $T(ax) = aT(x)$ *for all* $a \in \mathbb{Q}$. *[This does not imply* T *is a linear transformation, see Exercise 6.57.]*

EXERCISE 6.56. *(Removing the scaling property for the prime fields) Fix* p *as a prime number. Let* $\mathbb{F}_p = \{0, 1, ..., p - 1\}$ *be the finite field of size* p *(using mod* p *arithmetic). Let* U, V *be vector*

spaces over field $\mathbb{F}_p$. Let $T : U \to V$ be a function that satisfies $T(x + y) = T(x) + T(y)$ for all $x, y \in U$. In this special case, we show this information alone is enough to guarantee that T is a linear transformation.

a) Show that $T(0) = 0$.

b) Use the fact that if $a \in \{1, 2, ..., p - 1\}$, then $a = 1 + 1 + \cdots + 1$, to prove that $T(ax) = aT(x)$ for all $x \in U$ and $a \in \mathbb{F}_p$.

EXERCISE 6.57. *(The case $f(x + y) = f(x) + f(y)$) Let $f : \mathbb{R} \to \mathbb{R}$ be a discontinuous function that satisfies $f(x+y) = f(x)+f(y)$ for all $x, y \in \mathbb{R}$, but $f(cz) \neq cf(z)$ for some particular $z, c \in \mathbb{R}$. [Using a set theory axiom called the* axiom of choice*, it can be shown that such a function f exists.] Fix $n, m \in \mathbb{N}$. Define $T : \mathbb{R}^n \to \mathbb{R}^m$ by $T(x_1, ..., x_n) = f(x_1)$ for all $x = (x_1, ..., x_n) \in \mathbb{R}^n$. Show that $T(x + y) = T(x) + T(y)$ for all $x, y \in \mathbb{R}^n$, but $T(\theta u) = \theta T(u)$ can fail for some particular $\theta \in \mathbb{R}, u \in \mathbb{R}^n$.*

EXERCISE 6.58. *(The case $\mathbb{F}_4$) Let $\mathbb{F}_4 = \{0, 1, a, b\}$ be the finite field of size 4 from Exercise 1.17. [You will need the complete arithmetic tables for this exercise.] Build a function $T : \mathbb{F}_4 \to \mathbb{F}_4$ that satisfies $T(x + y) = T(x) + T(y)$ for all $x, y \in \mathbb{F}_4$, but $T(\theta u) = \theta T(u)$ fails for some $\theta, u \in \mathbb{F}_4$.*

EXERCISE 6.59. *(Polynomials with different degrees) Let V be the real vector space of functions $f : \mathbb{R} \to \mathbb{R}$. Let $\{p_1, p_2, ..., p_k\}$ be nonzero polynomial functions in V, so for $i \in \{1, ..., k\}$*

$$p_i(t) = a_{i,0} + a_{i,1}t + ... + a_{i,d[k]}t^{d[k]} \quad \forall t \in \mathbb{R}$$

where the $a_{i,j}$ coefficients are real numbers, $d[k]$ is the degree of polynomial p_k, and $a_{i,d[k]} \neq 0$. Suppose $d[1], ..., d[k]$ are distinct nonnegative integers. We want to show $\{p_1, p_2, ..., p_k\}$ is linearly independent.

a) Suppose $k = 1$, so there is one (nonzero) polynomial p_1 on the list. Prove the list is linearly independent.

b) Induction: Suppose that if our list has size k, then the list is linearly independent. (Part (a) shows this is true for $k = 1$.) Prove for $k + 1$. Hint: For simplicity, you can assume $d[1] < d[2] < \cdots < d[k] < d[k + 1]$. Remember to start your proof: "Suppose that $\sum_{i=1}^{k+1} \theta_i p_i = 0$ for some real numbers θ_i. We want to show..."

c) To understand why we needed to assume the degrees are distinct, give a simple counter-example when the degrees are not *distinct.*

EXERCISE 6.60. *(Proving vectors form a basis) Let V be a vector space over a field $\mathbb{F}$. Let S be a subspace of V. Suppose that $k, m \in \mathbb{N}$ and $x_1, ..., x_k, y_1, ..., y_n$ are vectors that satisfy*

- *$\{x_1, ..., x_k\}$ is a basis for S.*
- *$\{x_1, ..., x_k, y_1, ..., y_m\}$ is a basis for V.*

Let $q_1, ..., q_m$ be any vectors in S. Show $\{x_1, ..., x_k, y_1 - q_1, y_2 - q_2, ..., y_m - q_m\}$ is a basis for V.

EXERCISE 6.61. *(Matrix of T, G, and T + G) Let V be the real vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$ of degree ≤ 2. Define the functions $T : V \to V$ and $G : V \to V$ for $f \in V$ by*

$$T(f)(t) = f(t) + f'(t) \quad \forall t \in \mathbb{R}$$
$$G(f)(t) = 5f(3t - 8) \qquad \forall t \in \mathbb{R}$$

Define $H = T + G$. In the following, we consider matrices of linear transformations from V to V with respect to the basis $\{1, t, t^2\}$ for V.

a) Show that T is a linear transformation and find $Mat(T)$.

b) Show that G is a linear transformation and find $Mat(G)$.

c) Show that H is a linear transformation and find $Mat(H)$.

EXERCISE 6.62. *(Matrix of T, G, and G ∘ T) Let V be the real vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$ of degree ≤ 3. Define the functions $T : V \to V$ and $G : V \to V$ for $f \in V$ by*

$$T(f)(t) = f(t) + f''(t) \quad \forall t \in \mathbb{R}$$
$$G(f)(t) = f(t - 1) \qquad \forall t \in \mathbb{R}$$

Define $H = G \circ T$, so $H(f) = G(T(f))$ for $f \in V$. In the following, we consider matrices of linear transformations from V to V with respect to the basis $\{1, t, t^2, t^3\}$ for V.

a) Show that T is a linear transformation and find $Mat(T)$.

b) Show that G is a linear transformation and find $Mat(G)$.

c) Show that H is a linear transformation and find $Mat(H)$. Verify $Mat(H) = Mat(G)Mat(T)$.

d) Do the matrices $Mat(G)$ and $Mat(T)$ commute? Explain.

CHAPTER 7.

Playing with transposes and the function $x^\top A y$

Suppose we are given a matrix $A \sim m \times n$ and a column vector b of size m, both having entries in some field $\mathbb{F}$. We want to find a column vector $x \in \mathbb{F}^n$ that solves $Ax = b$. In general, this can be done by Gaussian elimination. However, Gaussian elimination does not yield a simple closed form answer. If A is square and invertible then we have a closed form solution $x^* = A^{-1}b$, and this solution is unique. However, if A is not square then A^{-1} *cannot* exist. Nevertheless, both $AA^\top$ and $A^\top A$ are square, and there is hope that at least one of these might be invertible.

LEMMA 7.1. *Let A be a $m \times n$ matrix with entries in some field $\mathbb{F}$. Fix $b \in \mathbb{F}^m$. Consider the problem of finding a vector $x \in \mathbb{F}^n$ that solves the equation $Ax = b$.*

a) If $AA^\top$ is invertible then one (possibly nonunique) solution to $Ax = b$ is

$$x^* = A^\top (AA^\top)^{-1} b$$

b) If $A^\top A$ is invertible, and if $Ax = b$ has at least one solution, then it has only one solution and that unique solution is

$$x^* = (A^\top A)^{-1} A^\top b$$

PROOF. Each part can be proven in just one or two lines (see Exercises 7.5 and 7.6). □

For example, for $\mathbb{F} = \mathbb{R}$ consider

$$\underbrace{\begin{bmatrix} 1 & 0 & -1 \\ 2 & 2 & 4 \end{bmatrix}}_{A}\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \underbrace{\begin{bmatrix} 8 \\ -1 \end{bmatrix}}_{b}$$

The matrix $AA^\top$ is

$$AA^\top = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 2 & 4 \end{bmatrix}\begin{bmatrix} 1 & 2 \\ 0 & 2 \\ -1 & 4 \end{bmatrix} = \begin{bmatrix} 2 & -2 \\ -2 & 24 \end{bmatrix}$$

Since $AA^\top$ is an invertible 2×2 matrix, part (a) of Lemma 7.1 implies that *one particular solution* to $Ax = b$ is

$$x^* = \begin{bmatrix} 1 & 2 \\ 0 & 2 \\ -1 & 4 \end{bmatrix}\begin{bmatrix} 2 & -2 \\ -2 & 24 \end{bmatrix}^{-1}\begin{bmatrix} 8 \\ -1 \end{bmatrix} = \begin{bmatrix} 4.954545454545454\ldots \\ 0.636363636363636\ldots \\ -3.045454545454545\ldots \end{bmatrix}$$

This is not the only solution to $Ax = b$ because $Null(A) \neq \{0\}$. Part (b) of Lemma 7.1 does not apply because the 3×3 matrix $A^\top A$ is not invertible.[1]

Lemma 7.1 can be viewed as a *precursor* to the theory of least squares developed in Chapter 9. Notes:

- There are cases when *neither* $A^\top A$ *nor* $AA^\top$ is invertible (such as when $A = 0$).
- $AA^\top$ can only be invertible if A is square or wide (see Exercise 7.5). If $AA^\top$ is invertible and A is wide, the vector $x^* = A^\top(AA^\top)^{-1}b$ from part (a) is not the only solution to $Ax = b$. However, if $\mathbb{F} = \mathbb{R}$, so all entries of A are real numbers, the theory of least squares can be used to show that $x^* = A^\top(AA^\top)^{-1}b$ is the solution that is *closest to the all-zero vector* (see Section 9.15). That is, it is the solution with minimum norm.
- $A^\top A$ can only be invertible if A is square or tall (see Exercise 7.6). Consider the special case when $A^\top A$ is invertible but $Ax = b$ does *not* have a solution. This can occur when A is tall (see also Exercise 7.7). Then the vector $x^* = (A^\top A)^{-1}A^\top b$ in part (b) of Lemma 7.1 is in a state of limbo: It exists, its structural form is the only possible solution to $Ax = b$, but it *cannot* be a solution because no solution exists. It is a *phantom solution*. We can

[1] If $A \sim 2 \times 3$ then $A^\top A$ has size 3×3 and *cannot* be invertible. This is because an invertible 3×3 matrix must have a 3-dimensional column space, but $Col(A^\top A) \subseteq Col(A^\top)$ and $Col(A^\top)$ has dimension at most 2 because $A^\top$ only has two columns.

define $b - Ax^*$ as the *error vector*. Under the field $\mathbb{F} = \mathbb{R}$, it turns out that $x^* = (A^\top A)^{-1} A^\top b$ is a unique *least squares* solution to $Ax \approx b$, meaning that $x^* \in \mathbb{R}^n$ minimizes $||b - Ax||^2$ over all other vectors $x \in \mathbb{R}^n$, where $||b - Ax||$ is the Euclidean norm of the error vector.

- If $\mathbb{F} = \mathbb{R}$, so the entries of A are real numbers, it can be shown that $A^\top A$ is invertible if and only if $Null(A) = \{0\}$ (see Exercise 7.7). Likewise, $AA^\top$ is invertible if and only if $Null(A^\top) = \{0\}$. This condition does not necessarily hold for a general field $\mathbb{F}$. For example, in the binary field $\mathbb{F}_2 = \{0, 1\}$ define the 2×1 matrix A as follows:

$$A = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies A^\top A = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = 1 + 1 = 0$$

  Then $Null(A) = \{0\}$ but $A^\top A$ is the 1×1 zero matrix, which is not invertible.

7.1. The function $x^\top A y$

Given any $m \times n$ matrix $A = (a_{ij})$, we can define a function $f : \mathbb{F}^m \times \mathbb{F}^n \to \mathbb{F}$ by

$$f(x, y) = x^\top A y \quad \forall x \in \mathbb{F}^m, y \in \mathbb{F}^n$$

This function is treated in Exercises 7.3, 7.9, 8.5. Notice that $x^\top A y$ can be viewed as the multiplication of three matrices of sizes $1 \times m$, $m \times n$, $n \times 1$, and hence its scalar output can be viewed as a 1×1 matrix.[2] If a is a 1×1 matrix (meaning a is a scalar in the field $\mathbb{F}$), then it is equal to its transpose. That is, $a \in \mathbb{F}$ implies $a = a^\top$. Therefore, given any vectors $x \in \mathbb{F}^m$ and $y \in \mathbb{F}^n$ we can define the scalar $a = x^\top A y$ and we conclude

$$x^\top A y = a = a^\top = y^\top A^\top x$$

The above equality needs $x^\top A y$ to be a scalar; For matrices A, B, C it does not generally hold that $(ABC)^\top = ABC$.

Suppose $x = (x_1, \ldots, x_m) \in \mathbb{F}^m$ and $y = (y_1, \ldots, y_n) \in \mathbb{F}^n$. The following follow from the definition of matrix multiplication:

- Given $i \in \{1, \ldots, m\}$, the ith row of the column vector Ay is equal to $\sum_{j=1}^n a_{ij} y_j$.
- Given $j \in \{1, \ldots, n\}$, the jth column of the row vector $x^\top A$ is equal to $\sum_{i=1}^m a_{ij} x_i$.

[2] Caveat: While it is sometimes useful to view a scalar as a 1×1 matrix, the operation of multiplying a scalar by a matrix can be done for a matrix of any size and hence does *not* follow the rules of multiplication by a 1×1 matrix.

LEMMA 7.2. *Given any $m \times n$ matrix $A = (a_{ij})$ with entries in some field $\mathbb{F}$, we have*

$$x^\top A y = \sum_{j=1}^{n} \sum_{i=1}^{m} x_i y_j a_{ij} \quad \forall x \in \mathbb{F}^m, y \in \mathbb{F}^n$$

PROOF. See Exercise 7.3. □

For example, take $\mathbb{F} = \mathbb{R}$ and

$$A = \begin{bmatrix} 2 & 0 \\ -1 & 1 \\ 0 & 1 \end{bmatrix}$$

Then for $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, $y = (y_1, y_2) \in \mathbb{R}^2$ we have

$$\begin{aligned} x^\top A y &= (2)x_1 y_1 + (-1)x_2 y_1 + (0)x_3 y_1 + (0)x_1 y_2 + (1)x_2 y_2 + (1)x_3 y_2 \\ &= 2x_1 y_1 - x_2 y_1 + x_2 y_2 + x_3 y_2 \end{aligned}$$

7.2. Positive definite and positive semidefinite

Here we consider the field $\mathbb{F} = \mathbb{R}$. This field is useful because any two real numbers x and y can be compared to determine whether or not $x \leq y$.

DEFINITION 7.1. *A square $n \times n$ matrix $A = (a_{ij})$ with real entries is* symmetric and positive semidefinite *if $A^\top = A$ and*

$$x^\top A x \geq 0 \quad \forall x \in \mathbb{R}^n$$

DEFINITION 7.2. *A square $n \times n$ matrix $A = (a_{ij})$ with real entries is* symmetric and positive definite *if $A^\top = A$ and*

$$x^\top A x \geq 0 \quad \forall x \in \mathbb{R}^n$$

with equality if and only if $x = 0$.

For example, consider

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

It is clear that A is symmetric. To show it is symmetric and positive definite, for all $x \in \mathbb{R}^2$

we have

$$x^\top Ax = \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 2 & -1 \\ -1 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2x_1^2 - 2x_1x_2 + 2x_2^2 = (x_1 - x_2)^2 + x_1^2 + x_2^2 \geq x_1^2 + x_2^2 \geq 0$$

This shows that A is symmetric and positive *semidefinite*. However, the equality $x_1^2 + x_2^2 = 0$ holds if and only if $x_1 = x_2 = 0$. So A is in fact symmetric and positive definite.

Related facts:

1. When A is a real and symmetric 2×2 matrix with the general form:

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

   - (A is positive semidefinite) $\iff$ $(a \geq 0, d \geq 0, ad \geq b^2)$ [See Exercise 7.10]
   - (A is positive definite) $\iff$ $(a > 0, d > 0, ad > b^2)$ [See Exercise 7.11]

2. It is possible for an *asymmetric* real $n \times n$ matrix A to satisfy $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^n$ (see Exercise 7.12).
3. If H is any real $m \times n$ matrix, Exercise 7.15 shows:
   - $H^\top H$ is a symmetric and positive semidefinite $n \times n$ matrix.
   - ($H^\top H$ is positive definite) $\iff$ $(Null(H) = \{0\})$ $\iff$ ($H^\top H$ is invertible)
4. Suppose A is a real $n \times n$ matrix. Then:
   - (A is symmetric and pos. semidefinite) $\iff$ ($A = H^\top H$ for some matrix H)
   - (A is symmetric and pos. definite) $\iff$ ($A = H^\top H$ for some invertible matrix H)

   The forward directions use the Spectral Theorem of Chapter 15. The reverse directions follow by the previous fact.

7.3. Exercises

EXERCISE 7.1. *Let A be a square matrix. Show that $A + A^\top$ is symmetric. Give two specific examples.*

EXERCISE 7.2. *Fix $\mathbb{F} = \mathbb{R}$. Give a specific example of a matrix A with at least two linearly independent columns. Compute $A^\top A$. Is $A^\top A$ invertible?*

EXERCISE 7.3. *(Expressing $x^\top Ay$) This exercise proves Lemma 7.2. Let $A = (a_{ij}) \sim m \times n$ with*

entries in some field $\mathbb{F}$. Fix $x \in \mathbb{F}^m$ and $y \in \mathbb{F}^n$. We want to show that

$$x^\top A y = \sum_{i=1}^{m} \sum_{j=1}^{n} x_i y_j a_{ij}$$

a) Define $w = Ay$. Show that w is a column vector of size m. Give a summation expression for w_i for each $i \in \{1, \ldots, m\}$.

b) Use part (a) to prove the desired result.

EXERCISE 7.4. *($A^\top A$ is symmetric and positive semidefinite) Suppose A is a $m \times n$ matrix with entries in some field $\mathbb{F}$.*

a) Show that $A^\top A$ is a symmetric $n \times n$ matrix.

b) Suppose $\mathbb{F} = \mathbb{R}$. Show that $A^\top A$ is a symmetric and positive semidefinite matrix. Hint: For a given $x \in \mathbb{R}^n$, define $w = Ax$ and consider $w^\top w$, which is the standard inner product of w with itself.

c) Explain why (a) and (b) imply that if $\mathbb{F} = \mathbb{R}$, then $AA^\top$ is a symmetric and positive semidefinite matrix of size $m \times m$.

EXERCISE 7.5. *(Proof of Lemma 7.1a) Fix $A \sim m \times n$ with entries in some field $\mathbb{F}$. Fix b as a $m \times 1$ column vector. We explore solutions to $Ax = b$.*

a) Suppose $AA^\top$ is invertible. Show that one solution to $Ax = b$ is $x^* = A^\top (AA^\top)^{-1} b$.

b) Show that $AA^\top$ cannot be invertible if A is tall. (Hint: Every wide matrix has a nontrivial null space.)

c) Give a specific example of a wide matrix A for which $AA^\top$ is invertible. Then find a b for which $Ax = b$ has more than one solution (including the solution proposed in part (a)).

EXERCISE 7.6. *(Proof of Lemma 7.1b) Fix $A \sim m \times n$ with entries in some field $\mathbb{F}$. Fix b as a $m \times 1$ column vector. We explore solutions to $Ax = b$.*

a) Suppose $A^\top A$ is invertible. Suppose $Ax = b$ has a solution. Show the one and only solution is $x^* = (A^\top A)^{-1} A^\top b$.

b) Show that $A^\top A$ cannot be invertible if A is wide. (Hint: Every wide matrix has a nontrivial null space.)

c) Give a specific example of a tall matrix A for which $A^\top A$ is invertible. Find a vector b for which $Ax = b$ has a solution. Then, find another vector b for which $Ax = b$ has no solution.

EXERCISE 7.7. *($Null(A) = Null(A^\top A)$) Consider the field $\mathbb{F} = \mathbb{R}$. Let $A \sim m \times n$ with entries in $\mathbb{R}$.*

a) Prove $Null(A) \subseteq Null(A^\top A)$.

b) Prove $Null(A^\top A) \subseteq Null(A)$. Hint: Fix $x \in Null(A^\top A)$. Consider $x^\top A^\top A x$ and the standard inner product for $\mathbb{R}^m$.

c) Show $A^\top A$ is invertible if and only if $Null(A) = \{0\}$. Explain why invertibility of $A^\top A$ implies A is square or tall.

d) Show that $Null(A) = Null(A^\top A)$ *fails in the binary field* $\mathbb{F}_2 = \{0, 1\}$ *with* $1 + 1 = 0$ *by considering the matrix*

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

EXERCISE 7.8. *(Constructing a matrix) This exercise relates to Lemma 7.1b and the discussion points given just after Lemma 7.1. Consider the field* $\mathbb{F} = \mathbb{R}$. *Construct a* 3×2 *matrix* A *such that* $A^\top A$ *is invertible. Find a vector* $b \in \mathbb{R}^3$ *for which* $Ax = b$ *has no solution. Compute the vector* $x^* = (A^\top A)^{-1} A^\top b$ *from Lemma 7.1b and show it does* not *solve* $Ax = b$. *Compute the error vector* $e = b - Ax^*$ *and its squared norm* $||e||^2$ *(using the standard Euclidean norm). Compute* $e^\top A$.

EXERCISE 7.9. *(Function* $f(x, y) = x^\top A y$*) Fix* $A \sim m \times n$ *with entries in a field* $\mathbb{F}$. *Let* $f : \mathbb{F}^m \times \mathbb{F}^n \to \mathbb{F}$ *be the function* $f(x, y) = x^\top A y$ *defined in Section 7.1.*

a) Suppose $n = m$. *Show that* $f(x, y) = f(y, x)$ *for all* $x, y \in \mathbb{F}^n$ *if and only if* A *is symmetric.*

b) Define the 3×2 *matrix:*

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

Express $f(x, y)$ *for a general* $x = (x_1, x_2, x_3) \in \mathbb{F}^3$ *and* $y = (y_1, y_2) \in \mathbb{F}^2$ *in the form of Lemma 7.2. That is, your expression should only involve the variables* x_1, x_2, x_3, y_1, y_2.

c) Consider part (b) in the binary field $\mathbb{F}_2 = \{0, 1\}$ *(so* $1 + 1 = 0$*). How large is the domain of* f*? Compute* $f(x, y)$ *for all* $x, y \in \{0, 1\}^3 \times \{0, 1\}^2$.

EXERCISE 7.10. *(*2×2 *positive semidefinite matrices) Consider the symmetric* 2×2 *matrix*

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

where $a, b, d \in \mathbb{R}$. *We want to show* A *is symmetric and positive semidefinite if and only if* $a \geq 0, d \geq 0, ad \geq b^2$.

a) Suppose $a \geq 0, d \geq 0, ad \geq b^2$. *Suppose either* $a = 0$ *or* $d = 0$. *Show that* A *is positive semidefinite.*

b) Suppose $a > 0$ *and* $d > 0$ *and* $ad \geq b^2$. *Show that* A *is positive semidefinite. Hint: Define* $\rho \in \mathbb{R}$ *so that* $\rho\sqrt{ad} = b$. *Then note* $(x\sqrt{a} + \rho y\sqrt{d})^2 = ax^2 + 2bxy + \rho^2 dy^2$.

c) Suppose A *is positive semidefinite. Show that* $a \geq 0$ *and* $d \geq 0$. *Hint: Consider* $x = (1, 0)$.

d) Suppose A *is positive semidefinite. Show that* $b^2 \leq ad$. *Hint: Consider* $(x, y) = (\sqrt{d}, \sqrt{a})$ *and* $(x, y) = (\sqrt{d}, -\sqrt{a})$.

EXERCISE 7.11. *(*2×2 *positive definite matrices) Extend the result and/or proof technique of Exercise 7.10 to show that the* 2×2 *matrix* A *considered there is symmetric and positive definite*

if and only if $a > 0, d > 0, ad > b^2$.

EXERCISE 7.12. *Give an example* 2×2 *matrix* A *with real entries such that* A *is* not *symmetric, but* $x^\top Ax \geq 0$ *for all* $x \in \mathbb{R}^2$. *Hint: Start with a symmetric positive definite matrix and then perturb at least one off-diagonal entry.*

EXERCISE 7.13. *Give a necessary and sufficient condition on the entries* $a, b, c, d \in \mathbb{R}$ *of a (possibly asymmetric)* 2×2 *matrix* A *for that matrix to satisfy* $x^\top Ax \geq 0$ *for all* $x \in \mathbb{R}^2$. *You can reuse the proof technique of Exercise 7.10.*

EXERCISE 7.14. *Let* A *be a* $n \times n$ *matrix with entries in* $\mathbb{R}$. *Suppose* $x^\top Ax \geq 0$ *for all* $x \in \mathbb{R}^n$.

a) Show that $x^\top A^\top x \geq 0$ *for all* $x \in \mathbb{R}^n$.

b) Show that $A + A^\top$ *is symmetric and positive semidefinite.*

EXERCISE 7.15. *(*$H^\top H$ *is positive semidefinite) Let* H *be a* $m \times n$ *matrix with entries in* $\mathbb{R}$. *It is clear that* $H^\top H$ *is symmetric. We want to show* $H^\top H$ *is positive semidefinite, and is positive definite if and only if* $Null(H) = \{0\}$.

a) Fix $x \in \mathbb{R}^n$. *Define* $w = Hx$. *Show that* $x^\top H^\top Hx = w^\top w$.

b) With $w = (w_1, \ldots, w_m) \in \mathbb{R}^m$, *show that* $w^\top w \geq 0$.

c) Show that if $Null(H) \neq 0$ *then* $H^\top H$ *is not positive definite.*

d) Show that if $Null(H) = \{0\}$ *then* $H^\top H$ *is positive definite. Hint: From part (a), if* $x^\top H^\top Hx = 0$ *then* $w^\top w = 0$.

EXERCISE 7.16. *Give an example real and symmetric* 3×3 *positive definite matrix with all entries nonzero.*

EXERCISE 7.17. *Give an example real and symmetric* 3×3 *positive definite matrix with exactly two zero entries.*

PART III.

Projections

CHAPTER 8.

Real inner product space

This chapter considers vector spaces V over the real field $\mathbb{R}$. The following additional structure is assumed: There is a function f that takes two vectors $x, y \in V$ as input, and outputs a real number $f(x, y)$ in such a way that three axioms of a *real inner product* are satisfied. This structure leads to a rich theory of orthogonality and projections.

8.1. Axioms for a real inner product

DEFINITION 8.1. *A function $f : V \times V \to \mathbb{R}$ is a* real inner product *on V if the following three properties hold:*

1. *Nonnegativity: For all $x \in V$ we have*

$$f(x, x) \geq 0$$

   Further, $f(x, x) = 0$ if and only if $x = 0$.

2. *Symmetry: For all $x, y \in V$ we have*

$$f(x, y) = f(y, x)$$

3. *Linearity in the second slot: For all $x, y, z \in V$ and all $a \in \mathbb{R}$ we have*

$$f(z, x + y) = f(z, x) + f(z, y)$$
$$f(z, ax) = af(z, x)$$

The above properties are called the three *axioms for a real inner product*. These properties are expressed above using standard function notation $f(x, y)$. However, inner products are usually expressed using notation $\langle x, y \rangle$. With this change in notation we have

$$\langle x, y \rangle = f(x, y) \quad \forall x, y \in V$$

Thus, a real inner product satisfies $\langle x, y \rangle \in \mathbb{R}$ for all vectors x, y, and

(i) Nonnegativity: $\langle x, x \rangle \geq 0$ for all $x \in V$, with equality if and only if $x = 0$.

(ii) Symmetry: $\langle x, y \rangle = \langle y, x \rangle$ for all $x, y \in V$.

(iii) Linearity in the second slot: For all $x, y, z \in V$ and $a \in \mathbb{R}$:

$$\langle z, x + y \rangle = \langle z, x \rangle + \langle z, y \rangle$$
$$\langle z, ax \rangle = a \langle z, x \rangle$$

The third property can be equivalently expressed:

$$\langle z, ax + by \rangle = a \langle z, x \rangle + b \langle z, y \rangle \quad \forall x, y, z \in V, \forall a, b \in \mathbb{R}$$

8.1.1. A quick note on complex inner products

A related definition of *complex inner product* is given in Chapter 10. It is worthwhile explaining the differences now:

- Complex inner products are used for vector spaces V over the field $\mathbb{C}$.
- The complex inner product $\langle x, y \rangle$ is allowed to take values in $\mathbb{C}$ (rather than always taking values in $\mathbb{R}$).
- The complex inner product $\langle x, y \rangle$ satisfies the same three axioms as a real inner product, with the exception that the symmetry axiom is replaced by *conjugate symmetry*: $\langle x, y \rangle = \overline{\langle y, x \rangle}$ for all $x, y \in V$ (where $\overline{z}$ denotes the complex conjugate of a scalar $z \in \mathbb{C}$).

We often just say "inner product" when the context is clear that we are using either a real vector space (so the inner product is real) or complex vector space (so the inner product is complex). This chapter considers only real vector spaces and real inner products.

8.1.2. Immediate consequences of the axioms

It is not difficult to show that every real inner product satisfies

$$\langle 0, x \rangle = \langle x, 0 \rangle = 0 \quad \forall x \in V$$

where 0 inside the inner product denotes the zero vector in V. Indeed, if $x \in V$ then

$$\begin{aligned}\langle 0, x\rangle &\overset{(a)}{=} \langle x, 0\rangle \\ &\overset{(b)}{=} \langle x, 0x\rangle \\ &\overset{(c)}{=} 0\,\langle x, x\rangle \\ &\overset{(d)}{=} 0\end{aligned}$$

where (a) uses symmetry of the inner product; (b) uses the fact

$$\underbrace{0}_{\in V} = \underbrace{0}_{\in \mathbb{R}} x$$

Also, (c) uses linearity in the second slot; (d) uses the fact that $\langle x, x\rangle$ is *some* real number, so it becomes 0 when multiplied by 0.

The symmetry property can be used to show that the second slot is not special: Every real inner product is also linear in the first slot. Indeed, for $x, y, z \in V$ and $a, b \in \mathbb{R}$ we have

$$\langle ax + by, z\rangle = \langle z, ax + by\rangle = a\,\langle z, x\rangle + b\,\langle z, y\rangle = a\,\langle x, z\rangle + b\,\langle y, z\rangle$$

Thus, $\langle x, y\rangle$ is a *bilinear function* (it is linear in y when x is fixed; linear in x when y is fixed). Therefore, for $x, y, w, z \in V$ and $a, b, c, d \in \mathbb{R}$ we have

$$\langle ax, by\rangle = b\,\langle ax, y\rangle = ab\,\langle x, y\rangle$$

$$\langle x + y, w + z\rangle = \langle x + y, w\rangle + \langle x + y, z\rangle = \langle x, w\rangle + \langle y, w\rangle + \langle x, z\rangle + \langle y, z\rangle$$

From this, it can be shown by induction that if $\{x_1, \ldots, x_m\}$ and $\{y_1, \ldots, y_n\}$ are any vectors, and if $a_1, \ldots, a_m$ and $b_1, \ldots, b_n$ are any real numbers, then (see Exercise 8.31):

$$\left\langle \sum_{i=1}^{m} a_i x_i, \sum_{j=1}^{n} b_j y_j \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j \left\langle x_i, y_j \right\rangle \tag{8.1}$$

A real vector space V can have many different inner products. For example, suppose $\langle x, y\rangle$ is a real inner product on V. Fix a as a positive real number and define $f(x, y) = a\,\langle x, y\rangle$ for all $x, y \in V$. It is not difficult to show $f(x, y)$ satisfies nonnegativity, symmetry, and linearity in the second slot. So $f(x, y)$ is another real inner product on V.

8.1.3. Terminology

A real vector space V that has a real inner product is called a *real inner product space*.

8.2. Cauchy-Schwarz inequality

LEMMA 8.1. *(Cauchy-Schwarz) Let $\langle x, y\rangle$ be a real inner product on a real vector space V. Then*

$$|\langle x, y\rangle| \leq \sqrt{\langle x, x\rangle \langle y, y\rangle} \quad \forall x, y \in V \tag{8.2}$$

Further, the inequality holds with equality if and only if either $y = 0$ or $x = ay$ for some $a \in \mathbb{R}$.

PROOF. We first prove the inequality (8.2). If $y = 0$ the inequality reduces to $0 \leq 0$, which is true. Suppose $y \neq 0$. Define

$$f(t) = \langle ty - x, ty - x\rangle \quad \forall t \in \mathbb{R}$$

We have $f(t) \geq 0$ for all $t \in \mathbb{R}$ (by the nonnegativity property of an inner product). However

$$f(t) = \langle ty - x, ty - x\rangle = t^2 \langle y, y\rangle - 2t \langle y, x\rangle + \langle x, x\rangle \quad \forall t \in \mathbb{R}$$

Since $y \neq 0$, property 1 of an inner product ensures $\langle y, y\rangle > 0$. So function $f(t)$ is a parabola and its minimizing point is the t for which $f'(t) = 0$:

$$f'(t) = 2t \langle y, y\rangle - 2 \langle y, x\rangle = 0$$

The minimizing point is $t^* = \frac{\langle y,x\rangle}{\langle y,y\rangle}$. Then

$$0 \leq f(t^*) = -\frac{\langle y, x\rangle^2}{\langle y, y\rangle} + \langle x, x\rangle$$

and so $\frac{\langle y,x\rangle^2}{\langle y,y\rangle} \leq \langle x, x\rangle$. Rearranging terms proves the desired inequality (8.2). Exercise 8.8 proves the necessary and sufficient condition for equality. □

8.3. Inner product norm

Let V be a real vector space. A *norm* on V is a function $||\cdot||$ that maps each element of V to a real number and satisfies the following three properties:

1. Nonnegativity: $||x|| \geq 0$ for all $x \in V$, with equality if and only if $x = 0$.
2. Scaling: $||ax|| = |a| \cdot ||x||$ for all $x \in V$ and $a \in \mathbb{R}$.
3. Triangle inequality: $||x + y|| \leq ||x|| + ||y||$ for all $x, y \in V$.

There can be many different norms on the same vector space. For example, scaling a norm by a positive real number creates a new norm. The next lemma shows that every inner product naturally creates a norm called the *inner product norm*.

LEMMA 8.2. *Let $\langle x, y\rangle$ be a real inner product on a real inner product space V. Define:*[1]

$$||x|| = \sqrt{\langle x, x\rangle} \quad \forall x \in V$$

Then $||x||$ is indeed a norm on V, and it is called the inner product norm.

PROOF. Fix $x \in V$ and $a \in \mathbb{R}$. We verify $||x||$ satisfies the three properties of a norm:

$$||x|| = \sqrt{\langle x, x\rangle} \geq 0 \tag{8.3}$$

$$||ax|| = \sqrt{\langle ax, ax\rangle} = \sqrt{a^2 \langle x, x\rangle} = |a| \cdot ||x|| \tag{8.4}$$

where (8.3) holds with equality if and only if $x = 0$ (by the first property of inner products). Equality (8.4) holds because the scalar a can be pulled out of both the first and second slots. Finally, Exercise 8.6 uses Cauchy-Schwarz to prove the triangle inequality. □

Unless otherwise stated, when working on a real inner product space with inner product $\langle x, y\rangle$, we use $||x||$ to denote the inner product norm $||x|| = \sqrt{\langle x, x\rangle}$. With this notation, the Cauchy-Schwarz inequality is

$$|\langle x, y\rangle| \leq ||x|| \cdot ||y|| \quad \forall x, y \in V$$

where equality holds if and only if either x is a scalar multiple of y, or y is a scalar multiple of x. The next lemma gives further properties of the inner product norm.

LEMMA 8.3. *(Properties of inner product norm) Let V be a real inner product space with inner product $\langle x, y\rangle$ and inner product norm $||x|| = \sqrt{\langle x, x\rangle}$. Then*

a) Norm-square of sum:

$$||x + y||^2 = ||x||^2 + ||y||^2 + 2\langle x, y\rangle \quad \forall x, y \in V$$

b) Parallelogram law:

$$||x + y||^2 + ||x - y||^2 = 2||x||^2 + 2||y||^2 \quad \forall x, y \in V$$

c) Polarization identity:

$$\langle x, y\rangle = \frac{||x + y||^2 - ||x - y||^2}{4} \quad \forall x, y \in V$$

[1] This definition $||x|| = \sqrt{\langle x, x\rangle}$ makes sense because the nonnegativity property of inner products ensures $\langle x, x\rangle \geq 0$ for all $x \in V$, so we can take square roots.

d) Strongly convex equality: For all $t \in \mathbb{R}$ we have

$$||tx + (1-t)y||^2 = t||x||^2 + (1-t)||y||^2 - t(1-t)||x-y||^2 \quad \forall x, y \in V$$

PROOF. Fix $x, y \in V$. Part (a) holds because

$$||x+y||^2 = \langle x+y, x+y\rangle = \langle x, x\rangle + \langle x, y\rangle + \langle y, x\rangle + \langle y, y\rangle$$

Each remaining part is proven by the previous parts. Details are left as an exercise. □

It can be shown that a norm $||\cdot||$ on a real vector space V is an inner product norm if and only if it satisfies the parallelogram law (Exercise 8.41). For example, the functions

$$||x||_1 = \sum_{i=1}^{n} |x_i| \quad \forall x \in \mathbb{R}^n$$
$$||x||_\infty = \max\{|x_1|, ..., |x_n|\} \quad \forall x \in \mathbb{R}^n$$

can be shown to be valid norms on $\mathbb{R}^n$ because they satisfy the nonnegativity, scaling, and triangle inequality properties. The first is called the L_1 *norm*, the second is the L_∞ *norm*. However, for $n \geq 2$, they both fail the parallelogram law, so neither is an inner product norm (consider $x = e_1$ and $y = e_2$). The strongly convex equality is a generalization of the parallelogram law: It reduces to the parallelogram law for $t = 1/2$. While it holds for all $t \in \mathbb{R}$, in optimization theory it is typically used for $t \in [0, 1]$ to conclude that the norm-square function satisfies a certain definition of a *strongly convex function*.[2]

8.4. Examples in $\mathbb{R}^n$

Fix n as a positive integer. Define the operation

$$\langle x, y\rangle = x^\top y \quad \forall x, y \in \mathbb{R}^n$$

It can be shown this is an inner product on $\mathbb{R}^n$. This is called the *standard inner product* on $\mathbb{R}^n$. The corresponding inner product norm is

$$||x|| = \sqrt{x^\top x} = \sqrt{\textstyle\sum_{i=1}^n x_i^2} \quad \forall x \in \mathbb{R}^n$$

This is the standard norm on $\mathbb{R}^n$, called the *Euclidean norm*, also called the L_2 norm. Unless otherwise stated, when working in the vector space $\mathbb{R}^n$ we implicitly use the standard inner product $x^\top y$ and its corresponding Euclidean norm. Using the standard inner product for

[2] Given $c > 0$, a function $f : V \to \mathbb{R}$ is said to be *c-strongly convex with respect to norm* $||\cdot||$ if $f(tx + (1-t)y) \leq tf(x) + (1-t)f(y) - \frac{ct(1-t)}{2}||x-y||^2$ for all $x, y \in V, t \in [0, 1]$.

$\mathbb{R}^n$, the Cauchy-Schwarz inequality becomes

$$|x^\top y| \le ||x|| \cdot ||y|| \quad \forall x, y \in \mathbb{R}^n$$

Equivalently,

$$\left|\sum_{i=1}^n x_i y_i\right| \le \sqrt{\sum_{i=1}^n x_i^2}\sqrt{\sum_{i=1}^n y_i^2} \quad \forall x, y \in \mathbb{R}^n \tag{8.5}$$

8.4.1. All inner products on $\mathbb{R}^n$

LEMMA 8.4. *A function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a real inner product on $\mathbb{R}^n$ if and only if*

$$f(x, y) = x^\top A y \quad \forall x, y \in \mathbb{R}^n$$

for some real $n \times n$ symmetric and positive definite matrix A. [Definition of "positive definite" is in Chapter 7.2.] The special case $A = I$ corresponds to the standard inner product on $\mathbb{R}^n$.

PROOF. See Exercise 8.5 for a proof of the reverse direction. To show the forward direction, suppose $f(x, y)$ is a real inner product on $\mathbb{R}^n$. A *Gram-Schmidt orthogonalization* result (Theorem 8.1 in Section 8.9) implies existence of a basis $\{x_1, ..., x_n\}$ for $\mathbb{R}^n$ that is *orthonormal* with respect to inner product f, meaning that

$$f(x_i, x_j) = 1_{\{i=j\}} \quad \forall i, j \in \{1, ..., n\} \tag{8.6}$$

where $1_{\{i=j\}}$ is the indicator function that is 1 if $i = j$, and 0 else. Let X be the $n \times n$ matrix with columns consisting of the column vectors $\{x_1, ..., x_n\}$. Since columns of X are linearly independent, matrix X is invertible. Define

$$A = (X^{-1})^\top X^{-1}$$

Exercise 8.37 shows that A is symmetric and positive definite. It remains to show $f(u, v) = u^\top A v$ for all $u, v \in \mathbb{R}^n$.

Let $\{e_1, ..., e_n\}$ be the standard basis for $\mathbb{R}^n$. Observe that

$$(X^{-1}X = I) \iff X^{-1}\begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \\ e_1 & e_2 & \cdots & e_n \\ | & | & \cdots & | \end{bmatrix}$$

Hence, for all $i \in \{1, ..., n\}$ we have $X^{-1}x_i = e_i$. Thus, for each $i, j \in \{1, ..., n\}$

$$x_i^\top A x_j \overset{(a)}{=} x_i^\top (X^{-1})^\top X^{-1} x_j = (X^{-1}x_i)^\top (X^{-1}x_j) \overset{(b)}{=} e_i^\top e_j \overset{(c)}{=} 1_{\{i=j\}}$$

where (a) substitutes the definition of A; (b) uses $X^{-1}x_i = e_i$ for all i; (c) uses $e_i^\top e_i = 1$ for $i \in \{1, \ldots, n\}$, and $e_i^\top e_j = 0$ for $i \neq j$. In view of (8.6), this means

$$x_i^\top A x_j = f(x_i, x_j) \quad \forall i, j \in \{1, \ldots, n\} \tag{8.7}$$

Now fix $u, v \in \mathbb{R}^n$. Let $(a_1, \ldots, a_n)$, $(b_1, \ldots, b_n)$ be the unique coefficients such that

$$u = \sum_{i=1}^n a_i x_i$$
$$v = \sum_{j=1}^n b_j x_j$$

Then

$$\begin{aligned} f(u, v) = f\left(\sum_{i=1}^n a_i x_i, \sum_{j=1}^n b_j x_j\right) &\overset{(a)}{=} \sum_{i=1}^n \sum_{j=1}^n a_i b_j f(x_i, x_j) \\ &\overset{(b)}{=} \sum_{i=1}^n \sum_{j=1}^n a_i b_j \left(x_i^\top A x_j\right) \\ &= \sum_{i=1}^n a_i x_i^\top \left(\sum_{j=1}^n b_j A x_j\right) \\ &= \left(\sum_{i=1}^n a_i x_i\right)^\top A \left(\sum_{j=1}^n b_j x_j\right) \\ &= u^\top A v \end{aligned}$$

where (a) uses the fact that f is a real inner product, hence, it is a bilinear function; (b) substitutes (8.7). □

8.4.2. Consequences for real matrices

LEMMA 8.5. *If A is a $m \times n$ matrix with real entries then:*

a) $A^\top A$ is a symmetric and positive semidefinite $n \times n$ matrix. [Definition of "positive semidefinite" is in Chapter 7.2.]

b) $Null(A^\top A) = Null(A)$ and

$$rank(A^\top A) = rank(A) = rank(A^\top) = rank(AA^\top)$$

PROOF. Note that $(A^\top A)^\top = A^\top A$, so $A^\top A$ is symmetric. For any $x \in \mathbb{R}^n$ we have

$$x^\top A^\top A x = (Ax)^\top (Ax) = ||Ax||^2 \geq 0$$

where we use the Euclidean norm. So $A^\top A$ is symmetric and positive semidefinite.

To prove (b), if $x \in Null(A)$ then $Ax = 0$ and so $A^\top Ax = 0$, so $x \in Null(A^\top A)$. Conversely, if $x \in Null(A^\top A)$ then

$$0 = A^\top Ax$$

Multiplying the above by $x^\top$ gives

$$0 = x^\top A^\top Ax = (Ax)^\top (Ax) = ||Ax||^2$$

Since $||Ax|| = 0$ we know $Ax = 0$, so $x \in Null(A)$. This completes the proof that $Null(A^\top A) = Null(A)$. Since A and $A^\top A$ both have n columns, the rank-nullity theorem implies

$$\text{rank}(A) + \dim Null(A) = n \tag{8.8}$$

$$\text{rank}(A^\top A) + \dim Null(A^\top A) = n \tag{8.9}$$

Since $Null(A) = Null(A^\top A)$, they have the same dimension, so (8.8)-(8.9) imply

$$\text{rank}(A) = \text{rank}(A^\top A)$$

This is true for any matrix with real entries, so it is true for the matrix $A^\top$, which means

$$\text{rank}(A^\top) = \text{rank}((A^\top)^\top A^\top) = \text{rank}(AA^\top)$$

We already know $\text{rank}(A) = \text{rank}(A^\top)$ by Lemma 5.4, which completes the chain of equalities for part (b). □

It is important to emphasize that Lemma 8.5 holds for *real matrices A* because $u^\top v$ is a *real inner product* for vectors $u, v \in \mathbb{R}^m$, which ensures:

$$\big((Ax)^\top (Ax) = 0\big) \iff (Ax = 0)$$

The result of Lemma 8.5 can fail if entries of A are in the binary field $\mathbb{F}_2 = \{0, 1\}$ because $u^\top v$ is not an inner product for $\mathbb{F}_2^m$ (there is no inner product for $\mathbb{F}_2^m$). Specifically,

$$\text{[In field } \mathbb{F}_2 = \{0,1\}\text{]:} \quad A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \implies A^\top A = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{rank}(A) = 1, \quad \text{rank}(A^\top A) = 0$$

See also Exercise 8.28 for a related example. Of course, Lemma 8.5 holds if we treat the

same matrix A in the real field:

$$\text{[In field } \mathbb{R}\text{]:} \quad A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \Longrightarrow A^\top A = \begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}, \quad \text{rank}(A) = \text{rank}(A^\top A) = 1$$

The result of Lemma 8.5 can also fail in the complex field $\mathbb{C}$ because (using $i^2 = -1$):

$$\text{[In field } \mathbb{C}\text{]:} \quad B = \begin{bmatrix} i & 0 \\ 1 & 0 \end{bmatrix} \Longrightarrow B^\top B = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{rank}(B) = 1, \quad \text{rank}(B^\top B) = 0$$

On the other hand, Chapter 10 defines a *complex inner product* on $\mathbb{C}^m$, from which a variation on Lemma 8.5 holds when $A^\top$ is replaced by A^H (where A^H is the *complex conjugate transpose* of A): $Null(A) = Null(A^H A)$ and $\text{rank}(A) = \text{rank}(A^H A)$. For example,

$$B = \begin{bmatrix} i & 0 \\ 1 & 0 \end{bmatrix} \Longrightarrow B^H B = \begin{bmatrix} -i & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} i & 0 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 0 \end{bmatrix}, \quad \text{rank}(B) = \text{rank}(B^H B) = 1$$

8.5. Examples in the space of square summable sequences

Let $\mathbb{R}^{\mathbb{N}}$ be the real vector space of infinite sequences of real numbers. So $x \in \mathbb{R}^{\mathbb{N}}$ if and only if

$$x = (x_1, x_2, x_3, \ldots) \text{ for some } x_i \in \mathbb{R} \text{ for } i \in \mathbb{N}$$

Define the following set $V \subseteq \mathbb{R}^{\mathbb{N}}$ of infinite real-valued sequences that are *square summable*:

$$V = \left\{ x \in \mathbb{R}^{\mathbb{N}} : \sum_{i=1}^{\infty} x_i^2 < \infty \right\}$$

It can be shown that V is a subspace of $\mathbb{R}^{\mathbb{N}}$. Therefore, V is itself a real vector space. The standard inner product on V is

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i \quad \forall x, y \in V$$

The reason that we have restricted attention to the set V of square summable sequences, rather than all real-valued sequences, is that the infinite sum $\sum_{i=1}^{\infty} x_i y_i$ can be shown to be a real number whenever $x, y \in V$. [To show this, use $-x_i^2 - y_i^2 \le 2x_i y_i \le x_i^2 + y_i^2$ for all i.] In contrast, the infinite sum may not be well defined when x and/or y is not square summable.

We now show that $\langle x, y \rangle$ is indeed an inner product:

1. Nonnegativity: For all $x \in V$ we have

$$\langle x, x \rangle = \sum_{i=1}^{\infty} x_i^2 \geq 0$$

Equality holds if and only if $x_i = 0$ for all $i \in \mathbb{N}$ because:

- If $x_i = 0$ for all $i \in \mathbb{N}$ then $\sum_{i=1}^{\infty} x_i^2 = \sum_{i=1}^{\infty} 0 = 0$.
- If $\sum_{i=1}^{\infty} x_i^2 = 0$ then for each $j \in \{1, 2, 3, \ldots\}$ we have $0 \leq x_j^2 \leq \sum_{i=1}^{\infty} x_i^2 = 0$, so $x_j = 0$.

2. Symmetry: For all $x, y \in V$ we have

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i = \sum_{i=1}^{\infty} y_i x_i = \langle y, x \rangle$$

3. Linearity in the second slot: For $x, y, z \in V$ and $a, b \in \mathbb{R}$ we have

$$\langle z, ax + by \rangle = \sum_{i=1}^{\infty} z_i (a x_i + b y_i) = a \sum_{i=1}^{\infty} z_i x_i + b \sum_{i=1}^{\infty} z_i y_i = a \langle z, x \rangle + b \langle z, y \rangle$$

Thus, $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$ is a valid inner product for V. With respect to this inner product, the Cauchy-Schwarz inequality is

$$\left| \sum_{i=1}^{\infty} x_i y_i \right| \leq \sqrt{\sum_{i=1}^{\infty} x_i^2} \sqrt{\sum_{i=1}^{\infty} y_i^2} \quad \forall x, y \in V \tag{8.10}$$

There are other inner products on V. Fix $\mu = (\mu_1, \mu_2, \ldots) \in \mathbb{R}^{\mathbb{N}}$ as any sequence of positive and bounded real numbers, so there is a value $\mu_{max} \in \mathbb{R}$ such that $0 < \mu_i \leq \mu_{max}$ for all $i \in \mathbb{N}$. It can be shown that the following is also an inner product on V:

$$\langle x, y \rangle = \sum_{i=1}^{\infty} \mu_i x_i y_i \quad \forall x, y \in V \tag{8.11}$$

The Cauchy-Schwarz inequality for this inner product is

$$\left| \sum_{i=1}^{\infty} \mu_i x_i y_i \right| \leq \sqrt{\sum_{i=1}^{\infty} \mu_i x_i^2} \sqrt{\sum_{i=1}^{\infty} \mu_i y_i^2} \quad \forall x, y \in V \tag{8.12}$$

Of course, (8.12) also follows directly from (8.10) by defining $\tilde{x} = (\sqrt{\mu_1} x_1, \sqrt{\mu_2} x_2, \ldots)$ and $\tilde{y} = (\sqrt{\mu_1} y_1, \sqrt{\mu_2} y_2, \ldots)$ and observing $(x, y \in V) \implies (\tilde{x}, \tilde{y} \in V)$, which is true more generally whenever $0 \leq \mu_i \leq \mu_{max}$ for all $i \in \mathbb{N}$. So if $\mu_i = 0$ for one or more indices $i \in \mathbb{N}$ then (8.12) is still true, but (8.11) fails to be an inner product on V.

8.6. Examples in the space of square integrable functions

Fix $a, b \in \mathbb{R}$ with $a < b$. Define V as the set of all functions $f : [a, b] \to \mathbb{R}$ that satisfy

$$\int_a^b f(t)^2 dt < \infty$$

These are called *square integrable functions*.[3] The set V is a vector space over the field $\mathbb{R}$ (see related Exercise 6.3). The standard inner product on this space is

$$\langle f, g \rangle = \int_a^b f(t)g(t)dt \quad \forall f, g \in V \tag{8.13}$$

Actually, this fails to be an inner product for a subtle reason described below, but we shall fix it!

First note that $\langle f, g \rangle$ indeed maps two square integrable functions $f, g \in V$ to a real number because

$$-\frac{f(t)^2}{2} - \frac{g(t)^2}{2} \le f(t)g(t) \le \frac{f(t)^2}{2} + \frac{g(t)^2}{2} \quad \forall t \in [a, b]$$

and so

$$-\infty < \int_a^b f(t)g(t)dt < \infty$$

It can further be shown that $\langle f, g \rangle$ satisfies the second and third properties required for an inner product (symmetry and linearity in the second slot). Finally, it satisfies the *first part* of the nonnegativity property: For any $f \in V$ we have

$$\langle f, f \rangle = \int_a^b f(t)^2 dt \ge 0$$

since the integral of a nonnegative function is nonnegative. *Unfortunately, equality does not hold if and only if $f(t)$ is the all-zero function*. Indeed, define the following function $z \in V$:

$$z(t) = \begin{cases} 42 & \text{if } t = (a+b)/2 \\ 0 & \text{else} \end{cases}$$

This function is not the all-zero function because $z((a+b)/2) = 42 \neq 0$. However, the function differs from the all-zero function only at one point, and that single point does

[3] Strictly speaking, one would also have to specify the integral used (Riemann, Lebesgue, etc.) and the mild additional assumptions on the function f to ensure it can be integrated. To be precise, as part of the *square integrable* definition we implicitly assume f is Borel measurable and the integral is the Lebesgue integral. All continuous functions are Borel measurable, as are all functions with at most finitely many discontinuities.

not affect the value of any integral. So

$$\langle z, z\rangle = \int_a^b z(t)^2 dt = 0 \quad \text{(but } z \neq 0\text{)}$$

Strictly speaking, this means that $\langle f, g\rangle = \int_a^b f(t)g(t)dt$ is *not* a real inner product on V.

We now fix it: We turn $\langle f, g\rangle$ into a real inner product through a standard technique of defining *equivalence classes* over the functions in V: Two functions $f, g \in V$ are said to be *equivalent*, written $f \equiv g$, if

$$\int_a^b (f(t) - g(t))^2 dt = 0$$

It follows that the $z(t)$ function defined above is *equivalent to the always-zero function*. It can be shown that if $f \equiv \tilde{f}$ and $g \equiv \tilde{g}$ then

$$\int_a^b f(t)g(t)dt = \int_a^b \tilde{f}(t)\tilde{g}(t)dt$$

It can be shown that two functions in V that differ only on a finite or countably infinite number of points in $[a, b]$ are equivalent.[4] If we redefine the vectors in V according to the equivalence classes, so all equivalent functions are treated as the same vector, then $\int_a^b f(t)g(t)dt$ is an inner product on V. It satisfies $\langle f, f\rangle = 0$ if and only if $f \equiv 0$ because

$$\langle f, f\rangle = 0 \iff \int_a^b f(t)^2 dt = 0 \iff \int_a^b (f(t) - 0)^2 dt = 0 \iff f \equiv 0$$

The inner product norm for this inner product is

$$\|f\| = \sqrt{\int_a^b f(t)^2 dt} \quad \forall f \in V$$

With respect to this inner product, the Cauchy-Schwarz inequality is

$$\left|\int_a^b f(t)g(t)dt\right| \leq \sqrt{\int_a^b f(t)^2 dt}\sqrt{\int_a^b g(t)^2 dt} \quad \forall f, g \in V$$

The concept of *equivalence class* is a technical detail that readers should not obsess over. It is common to refer to (8.13) as the standard inner product for real-valued square integrable functions (the equivalence class concept is often unstated but is used implicitly).

[4] From a measure theory perspective, if $f : [a, b] \to \mathbb{R}$ and $g : [a, b] \to \mathbb{R}$ are Borel measurable functions then: (i) $\{t \in [a, b] : f(t) \neq g(t)\}$ is a Borel measurable subset of $[a, b]$; (ii) $\int_a^b (f(t) - g(t))^2 dt = 0$ if and only if $\{t \in [a, b] : f(t) \neq g(t)\}$ is a subset of $[a, b]$ that has measure zero.

Another real inner product on V (using the same definition of equivalence class) is

$$\langle f, g\rangle = \int_a^b e^t f(t)g(t)dt, \quad \|f\| = \sqrt{\langle f, f\rangle} = \sqrt{\int_a^b e^t f(t)^2 dt}$$

Indeed, this can be shown to be finite whenever $f, g \in V$, and it can be shown to satisfy the nonnegativity, symmetry, and linearity in the second slot properties.

A useful workaround: To avoid equivalence classes, it is often useful to work in the real vector space $\tilde{V}$ of all *continuous* functions $f : [a, b] \to \mathbb{R}$. The only continuous function that is equivalent to 0 is the always-zero function. So, two continuous functions are equivalent if and only if they are always-equal. Also, when we use a closed and bounded interval $[a, b]$ for the domain, all continuous functions $f : [a, b] \to \mathbb{R}$ are automatically square integrable. This is because: (i) Continuous functions are Borel measurable; (ii) If $f(t)$ is continuous over $[a, b]$ then so is $f(t)^2$, and the extreme value theorem ensures there is a real number β such that

$$f(t)^2 \leq \beta \quad \forall t \in [a, b]$$

and so

$$\int_a^b f(t)^2 dt \leq \beta^2(a - b) < \infty$$

8.7. Example in the space of random variables with finite second moments

Let $(\Omega, \mathcal{F}, P)$ be a probability space, so Ω is a nonempty set (called the *sample space*); $\mathcal{F}$ is a sigma algebra on Ω (called the set of *events*); $P : \mathcal{F} \to [0, 1]$ is a probability measure. Define V as the real vector space of random variables $X : \Omega \to \mathbb{R}$ with finite second moment, so

$$\mathbb{E}\left[X^2\right] < \infty$$

Define two random variables $X, Y \in V$ to be *equivalent*, denoted $X \equiv Y$, if $P[X = Y] = 1$. So two random variables are equivalent if they are *almost surely equal*. Equivalent random variables are treated as the same vector in V. Using this idea of equivalence, it can be shown that the following is a real inner product on V:[5]

$$\langle X, Y\rangle = \mathbb{E}[XY] \quad \forall X, Y \in V$$

With this inner product, the Cauchy-Schwarz inequality becomes

$$|\mathbb{E}[XY]| \leq \sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]} \quad \forall X, Y \in V \tag{8.14}$$

[5] If X and Y are random variables then $(P[X = Y] = 1) \iff \mathbb{E}\left[(X - Y)^2\right] = 0$. So $\mathbb{E}\left[(X - 0)^2\right] = 0$ if and only if $X \equiv 0$. Let $U \sim Unif[0, 1]$ and assume $\{U = 0.7\}$ is nonempty. Define $W = U1_{\{U \neq 0.7\}}$. Then $W \equiv U$ but $W \neq U$.

For $X, Y \in V$ define the *covariance*, *variance*, and *correlation coefficient* by

$$\begin{aligned} Cov(X,Y) &= \mathbb{E}\left[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])\right] \\ Var(X) &= \mathbb{E}\left[(X - \mathbb{E}[X])^2\right] \\ \rho_{X,Y} &= \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}} \end{aligned}$$

where $\rho_{X,Y}$ is only defined when $Var(X) > 0$ and $Var(Y) > 0$. By defining $\tilde{X} = X - \mathbb{E}[X]$ and $\tilde{Y} = Y - \mathbb{E}[Y]$ and observing that $\tilde{X}, \tilde{Y} \in V$, we obtain

$$\begin{aligned} \langle \tilde{X}, \tilde{Y} \rangle &= \mathbb{E}\left[\tilde{X}\tilde{Y}\right] = Cov(X,Y) \\ \langle \tilde{X}, \tilde{X} \rangle &= \mathbb{E}\left[\tilde{X}^2\right] = Var(X) \end{aligned}$$

Therefore, applying the Cauchy-Schwarz inequality (8.14) to $\tilde{X}, \tilde{Y}$ gives

$$|Cov(X,Y)| \leq \sqrt{Var(X)Var(Y)} \quad \forall X, Y \in V$$

and so, assuming $Var(X) > 0$ and $Var(Y) > 0$, we obtain

$$-1 \leq \rho_{X,Y} \leq 1$$

The covariance $Cov(X,Y)$ is itself a function from $V \times V$ to $\mathbb{R}$ that satisfies symmetry and linearity in the second slot (hence, also linearity in the first slot). Also, $Cov(X,X) \geq 0$ for all $X \in V$. However, it does not satisfy the nonnegativity property of an inner product because, considering the constant random variable $Y(\omega) = 5$ for all $\omega \in \Omega$, we have $Cov(Y,Y) = 0$ (and Y is not equivalent to 0).

8.8. Orthogonal vectors

Let V be a real inner product space with inner product $\langle x, y \rangle$.

DEFINITION 8.2. *Two vectors $x, y \in V$ are said to be* orthogonal *if*

$$\langle x, y \rangle = 0$$

If x and y are orthogonal, then any scalar multiples of x and y are also orthogonal. Also, it is not difficult to show that the zero vector in V is orthogonal to all other vectors in V, that is, $\langle 0, x \rangle = 0$ for all $x \in V$.

LEMMA 8.6. *Let V be a real inner product space.*

a) If $x, y \in V$ are orthogonal then

$$||x + y||^2 = ||x||^2 + ||y||^2$$

b) If $x_1, \ldots, x_n$ are vectors in V that are pairwise orthogonal, so that $\langle x_i, x_j \rangle = 0$ for all $i \neq j$, then

$$\left|\left|\sum_{i=1}^n x_i\right|\right|^2 = \sum_{i=1}^n ||x_i||^2$$

PROOF. To prove part (a), fix $x, y \in V$. If x and y are orthogonal then

$$||x + y||^2 = ||x||^2 + ||y||^2 + 2\underbrace{\langle x, y \rangle}_{0}$$

To prove part (b), suppose $\{x_1, \ldots, x_n\}$ are pairwise orthogonal. Then

$$\begin{aligned}\left|\left|\sum_{i=1}^n x_i\right|\right|^2 &= \left\langle \sum_{i=1}^n x_i, \sum_{j=1}^n x_j \right\rangle \\ &= \sum_{i=1}^n \sum_{j=1}^n \langle x_i, x_j \rangle \\ &= \sum_{i=1}^n \underbrace{\langle x_i, x_i \rangle}_{||x_i||^2} + \sum_{i \neq j} \underbrace{\langle x_i, x_j \rangle}_{0}\end{aligned}$$

□

LEMMA 8.7. *If vectors $\{x_1, \ldots, x_n\}$ are pairwise orthogonal, and if none are zero, then they are linearly independent.*

PROOF. Suppose

$$\sum_{i=1}^n a_i x_i = 0$$

for some scalars $a_i \in \mathbb{R}$. We want to show $a_i = 0$ for all i. Fix $j \in \{1, \ldots, n\}$. Taking an inner product of both sides of the above equation with x_j gives

$$\left\langle \sum_{i=1}^n a_i x_i, x_j \right\rangle = 0$$

Using the property of linearity in the first slot gives

$$\sum_{i=1}^n a_i \langle x_i, x_j \rangle = 0$$

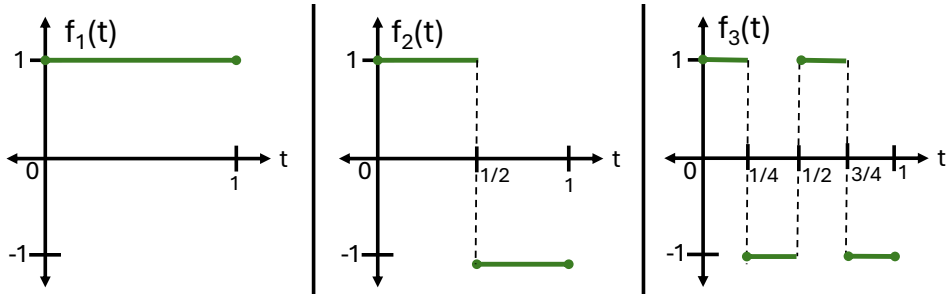


FIGURE 8.1. An illustration of the orthonormal functions $\{f_1, f_2, f_3\}$.

Since $\langle x_i, x_j \rangle = 0$ for all $i \neq j$, the above reduces to

$$a_j \langle x_j, x_j \rangle = 0$$

and so $a_j \left\| x_j \right\|^2 = 0$. Since $x_j \neq 0$ (hence $\left\| x_j \right\|^2 \neq 0$), we know $a_j = 0$. □

8.8.1. Example of orthogonal functions

Let V be the real vector space of square-integrable functions $f : [0,1] \to \mathbb{R}$. Define $f_1, f_2, f_3 \in V$ by (see Fig. 8.1):

$$f_1(t) = 1 \quad \forall t \in [0,1]$$

$$f_2(t) = \begin{cases} 1 & \text{if } 0 \leq t < 1/2 \\ -1 & \text{if } 1/2 \leq t \leq 1 \end{cases}$$

$$f_3(t) = \begin{cases} 1 & \text{if } t \in [0,1/4) \cup [1/2,3/4) \\ -1 & \text{if } t \in [1/4,1/2) \cup [3/4,1] \end{cases}$$

These functions are defined to have unit energy (norm-squared) and to be pairwise orthogonal:

$$\|f_1\|^2 = \int_0^1 f_1(t)^2 dt = \int_0^1 1 dt = 1$$

$$\|f_2\|^2 = \int_0^1 f_2(t)^2 dt = \int_0^1 1 dt = 1$$

$$\left\|f_3\right\|^2 = \int_0^1 f_3(t)^2 dt = \int_0^1 1 dt = 1$$

$$\langle f_1, f_2 \rangle = \int_0^1 f_1(t) f_2(t) dt = \int_0^{1/2} 1 dt - \int_{1/2}^1 1 dt = 0$$

$$\langle f_1, f_3 \rangle = \int_0^1 f_1(t) f_3(t) dt = \int_0^{1/4} 1 dt - \int_{1/4}^{1/2} 1 dt + \int_{1/2}^{3/4} 1 dt - \int_{3/4}^1 1 dt = 0$$

$$\langle f_2, f_3 \rangle = \int_0^1 f_2(t) f_3(t) dt = \int_0^{1/4} 1 dt - \int_{1/4}^{1/2} 1 dt - \int_{1/2}^{3/4} 1 dt + \int_{3/4}^1 1 dt = 0$$

Fix $a_1, a_2, a_3 \in \mathbb{R}$. The energy in $a_1 f_1 + a_2 f_2 + a_3 f_3$ is

$$\left\| a_1 f_1 + a_2 f_2 + a_3 f_3 \right\|^2 = \int_0^1 (a_1 f_1(t) + a_2 f_2(t) + a_3 f_3(t))^2 dt$$

Since f_1, f_2, f_3 are pairwise orthogonal, $a_1 f_1, a_2 f_2, a_3 f_3$ are also pairwise orthogonal, so Lemma 8.6 implies

$$\begin{aligned} \left\| a_1 f_1 + a_2 f_2 + a_3 f_3 \right\|^2 &= \|a_1 f_1\|^2 + \|a_2 f_2\|^2 + \left\| a_3 f_3 \right\|^2 \\ &= a_1^2 \|f_1\|^2 + a_2^2 \|f_2\|^2 + a_3^2 \left\| f_3 \right\|^2 \\ &= a_1^2 + a_2^2 + a_3^2 \end{aligned}$$

Lemma 8.7 implies the functions $\{f_1, f_2, f_3\}$ are linearly independent.

8.8.2. Example of orthogonal vectors in $\mathbb{R}^3$

Let $x_1 = (0, 1/\sqrt{2}, 1/\sqrt{2})$ and $x_2 = (\sqrt{1/5}, \sqrt{2/5}, -\sqrt{2/5})$. These are vectors in $\mathbb{R}^3$ that have norm 1 and are orthogonal (using the standard inner product on $\mathbb{R}^3$):

$$\begin{aligned} \|x_1\|^2 &= 0^2 + 1/2 + 1/2 = 1 \\ \|x_2\|^2 &= 1/5 + 2/5 + 2/5 = 1 \\ x_1^\top x_2 &= (0)\sqrt{1/5} + (1/\sqrt{2})\sqrt{2/5} - (1/\sqrt{2})\sqrt{2/5} = 0 \end{aligned}$$

Exercise: Find as many additional norm-1 vectors $\{x_3, \ldots, x_k\}$ as possible such that $\{x_1, \ldots, x_k\}$ are pairwise orthogonal.

Solution: If $\{x_1, \ldots, x_k\}$ are pairwise orthogonal in $\mathbb{R}^3$, and none are zero, they must be linearly independent (Lemma 8.7). Since we cannot have a collection of more than three linearly independent vectors in $\mathbb{R}^3$, at most we can find one more vector x_3. Let us construct $x_3 = (a, b, c)$ using the fact that we need $x_1^\top x_3 = 0$ and $x_2^\top x_3 = 0$:

$$\underbrace{\begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ \sqrt{1/5} & \sqrt{2/5} & -\sqrt{2/5} \end{bmatrix}}_{A} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

By Gaussian elimination we have

$$\begin{bmatrix} 0 & 1/\sqrt{2} & 1/\sqrt{2} \\ \sqrt{1/5} & \sqrt{2/5} & -\sqrt{2/5} \end{bmatrix} \to \begin{bmatrix} \sqrt{1/5} & \sqrt{2/5} & -\sqrt{2/5} \\ 0 & 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \to \begin{bmatrix} 1 & \sqrt{2} & -\sqrt{2} \\ 0 & 1 & 1 \end{bmatrix}$$
$$\to \begin{bmatrix} \boxed{1} & 0 & -2\sqrt{2} \\ 0 & \boxed{1} & 1 \end{bmatrix}$$

There is one free variable (variable c):

$$a - (2\sqrt{2})c = 0 \implies a = (2\sqrt{2})c$$
$$b + c = 0 \implies b = -c$$

Thus, all possible solutions are

$$Null(A) = \{((2\sqrt{2})c, -c, c) \in \mathbb{R}^3 : c \in \mathbb{R}\}$$

We need to choose c to make $((2\sqrt{2})c, -c, c)$ have norm 1:

$$1 = ||(a, b, c)||^2 = ((2\sqrt{2})c)^2 + c^2 + c^2 = 10c^2$$

There are two possibilities: $c = 1/\sqrt{10}$ and $c = -1/\sqrt{10}$. Let us choose the first. Then

$$\boxed{x_3 = (a, b, c) = (2/\sqrt{5}, -1/\sqrt{10}, 1/\sqrt{10})}$$

We should verify this is correct:

$$||x_3||^2 = 4/5 + 1/10 + 1/10 = 1$$
$$x_1^\top x_3 = (0)(2/\sqrt{5}) + (1/\sqrt{2})(-1/\sqrt{10}) + (1/\sqrt{2})(1/\sqrt{10}) = 0$$
$$x_2^\top x_3 = \sqrt{1/5}(2/\sqrt{5}) + \sqrt{2/5}(-1/\sqrt{10}) - \sqrt{2/5}(1/\sqrt{10}) = 0$$

8.9. Gram-Schmidt procedure

Let V be a real inner product space. A collection of vectors $\{v_1, ..., v_n\}$ in V is said to be *orthonormal* if

- They are pairwise orthogonal, so $\left\langle v_i, v_j \right\rangle = 0$ for all $i \neq j$.
- They are normal, so $||v_i|| = 1$ for all $i \in \{1, ..., n\}$.

Orthonormal vectors are orthogonal and nonzero, so they must be linearly independent. If they form a basis for some subspace S of V, they are called an *orthonormal basis*.

THEOREM 8.1. *(Gram-Schmidt) If V is a real inner product space and S is a nontrivial and finite-dimensional subspace of V, then S has an orthonormal basis.*

The proof of this theorem is constructive: Given any basis $\{v_1, ..., v_n\}$ for S we construct an orthonormal basis for S according to the following *Gram-Schmidt procedure*:

- **Step 1:** We know $v_1 \neq 0$ because it is part of the basis. Define

$$q_1 = \frac{v_1}{||v_1||} \tag{8.15}$$

  Then $||q_1|| = 1$ (q_1 is the normalized version of v_1). Note that

$$Span\{q_1\} = Span\{v_1\} \tag{8.16}$$

- **Step 2:** Define

$$Q_2 = v_2 - \langle v_2, q_1 \rangle q_1 \tag{8.17}$$

$$q_2 = \frac{Q_2}{||Q_2||} \tag{8.18}$$

  This avoids division by zero because $Q_2 \neq 0$; that is because $\{v_1, v_2\}$ are linearly independent and so

$$v_2 \notin Span\{v_1\} = Span\{q_1\}$$

  which uses (8.16). So $||q_2|| = 1$. Notice that Q_2 is orthogonal to q_1:

$$\langle Q_2, q_1 \rangle = \langle v_2 - \langle v_2, q_1 \rangle q_1, q_1 \rangle = \langle v_2, q_1 \rangle - \langle v_2, q_1 \rangle \underbrace{\langle q_1, q_1 \rangle}_{1} = 0$$

  Thus, q_2 is also orthogonal to q_1. So $\{q_1, q_2\}$ are orthonormal. We now claim:

$$Span\{q_1, q_2\} = Span\{v_1, v_2\} \tag{8.19}$$

  To prove (8.19), we show for each $i \in \{1, 2\}$ that $q_i \in Span\{v_1, v_2\}$ and $v_i \in Span\{q_1, q_2\}$. Equation (8.16) already implies

$$q_1 \in Span\{v_1\} \subseteq Span\{v_1, v_2\}$$
$$v_1 \in Span\{q_1\} \subseteq Span\{q_1, q_2\}$$

It remains only to prove $v_2 \in Span\{q_1, q_2\}$ and $q_2 \in Span\{v_1, v_2\}$. From (8.17) we obtain

$$v_2 = Q_2 + \langle v_2, q_1 \rangle q_1 \overset{(a)}{=} \underbrace{\|Q_2\|}_{\text{scalar}} q_2 + \underbrace{\langle v_2, q_1 \rangle}_{\text{scalar}} q_1 \tag{8.20}$$

where equality (a) holds by substituting (8.18). Thus $v_2 \in Span\{q_1, q_2\}$. Also, rearranging equation (8.20) gives

$$q_2 = \frac{1}{\|Q_2\|} v_2 - \frac{\langle v_2, q_1 \rangle}{\|Q_2\|} q_1 \overset{(a)}{=} \underbrace{\frac{1}{\|Q_2\|}}_{\text{scalar}} v_2 - \underbrace{\frac{\langle v_2, q_1 \rangle}{\|Q_2\|} \frac{1}{\|v_1\|}}_{\text{scalar}} v_1$$

where (a) holds by (8.15). Thus, $q_2 \in Span\{v_1, v_2\}$, completing the proof of (8.19).

- **General step** $k \in \{3, ..., n\}$**:** Assume from previous steps that we have manufactured orthonormal vectors $\{q_1, ..., q_{k-1}\}$ such that

$$Span\{q_1, ..., q_{k-1}\} = Span\{v_1, ..., v_{k-1}\} \tag{8.21}$$

(This is indeed true for $k = 3$ by (8.19).) Define

$$Q_k = v_k - \sum_{i=1}^{k-1} \langle v_k, q_i \rangle q_i \tag{8.22}$$

$$q_k = \frac{Q_k}{\|Q_k\|} \tag{8.23}$$

This avoids division by zero since $Q_k \neq 0$; that is because $\{v_1, ..., v_k\}$ are linearly independent so

$$v_k \notin Span\{v_1, ..., v_{k-1}\} = Span\{q_1, ..., q_{k-1}\}$$

which uses (8.21). Q_k is orthogonal to $q_1, q_2, ..., q_{k-1}$ because for each $j \in \{1, ..., k-1\}$:

$$\left\langle Q_k, q_j \right\rangle = \left\langle v_k - \sum_{i=1}^{k-1} \langle v_k, q_i \rangle q_i, q_j \right\rangle = \left\langle v_k, q_j \right\rangle - \sum_{i=1}^{k-1} \langle v_k, q_i \rangle \underbrace{\left\langle q_i, q_j \right\rangle}_{1_{\{i=j\}}} = \left\langle v_k, q_j \right\rangle - \left\langle v_k, q_j \right\rangle = 0$$

which uses the fact $\left\langle q_i, q_j \right\rangle = 1_{\{i=j\}}$, which is an indicator function that is 1 if $i = j$, and 0 if $i \neq j$. Thus, $\{q_1, ..., q_{k-1}, q_k\}$ are orthonormal. We now claim:

$$Span\{q_1, ..., q_k\} = Span\{v_1, ..., v_k\} \tag{8.24}$$

To prove (8.24), we want to show for each $i \in \{1, ..., k\}$ that $q_i \in Span\{v_1, ..., v_k\}$ and

$v_i \in Span\{q_1, ..., q_k\}$. We already know (8.21), so it suffices to prove $v_k \in Span\{q_1, ..., q_k\}$ and $q_k \in Span\{v_1, ..., v_k\}$. From (8.22) we obtain

$$v_k = Q_k + \sum_{i=1}^{k-1} \langle v_k, q_i \rangle q_i \overset{(a)}{=} ||Q_k|| q_k + \sum_{i=1}^{k-1} \langle v_k, q_i \rangle q_i$$

where equality (a) uses (8.23). Thus, $v_k \in Span\{q_1, ..., q_k\}$. Finally, by (8.22)-(8.23)

$$q_k = \frac{v_k}{||Q_k||} - \sum_{i=1}^{k-1} \frac{\langle v_k, q_i \rangle}{||Q_k||} q_i \tag{8.25}$$

and

$$-\sum_{i=1}^{k-1} \frac{\langle v_k, q_i \rangle}{||Q_k||} q_i \in Span\{q_1, ..., q_{k-1}\} = Span\{v_1, ..., v_{k-1}\}$$

which uses (8.21). Hence, (8.25) implies $q_k \in Span\{v_1, ..., v_k\}$. □

8.9.1. Relation to projection theory

In the language of projection theory (Chapter 9), the general step (8.22) of the Gram-Schmidt procedure can be understood as follows: Define

$$p = \sum_{i=1}^{k-1} \langle v_k, q_i \rangle q_i$$

Then p is the *projection of vector v_k onto the subspace $Span\{q_1, ..., q_{k-1}\}$*. The projection theorem of Chapter 9 ensures the error vector $v_k - p$ is orthogonal to every vector in $Span\{q_1, ..., q_{k-1}\}$. So we define $Q_k = v_k - p$ as this error vector.

8.9.2. Example: Gram-Schmidt for vectors in $\mathbb{R}^4$

Problem: Define the following three dimensional subspace of $\mathbb{R}^4$:

$$S = Span\{(1, 1, 0, 0), (1, 2, 3, 0), (1, 1, 1, 1)\}$$

Use the Gram-Schmidt procedure to construct an orthonormal basis for S (using the standard inner product for $\mathbb{R}^4$).

Solution: Define

$$\boxed{q_1 = \frac{(1, 1, 0, 0)}{||(1, 1, 0, 0)||} = \frac{(1, 1, 0, 0)}{\sqrt{2}}}$$

Define

$$\begin{aligned} Q_2 &= (1,2,3,0) - \langle (1,2,3,0), q_1 \rangle \, q_1 \\ &= (1,2,3,0) - \left\langle (1,2,3,0), \frac{(1,1,0,0)}{\sqrt{2}} \right\rangle \frac{(1,1,0,0)}{\sqrt{2}} \\ &= (1,2,3,0) - \frac{3}{2}(1,1,0,0) \\ &= (-1/2, 1/2, 3, 0) \end{aligned}$$

$$\boxed{q_2 = \frac{Q_2}{||Q_2||} = \frac{(-1/2,1/2,3,0)}{||(-1/2,1/2,3,0)||} = \frac{(-1/2,1/2,3,0)}{\sqrt{19/2}}}$$

As a preliminary check for errors, we have:

$$\begin{aligned} ||q_1||^2 &= 1/2 + 1/2 = 1 \\ ||q_2||^2 &= (1/4)/(19/2) + (1/4)/(19/2) + 9/(19/2) = 1 \\ q_1^\top q_2 &= \frac{1}{\sqrt{2}\sqrt{19/2}}(-1/2 + 1/2) = 0 \end{aligned}$$

Proceeding with the Gram-Schmidt procedure gives

$$\begin{aligned} Q_3 &= (1,1,1,1) - \langle (1,1,1,1), q_1 \rangle \, q_1 - \langle (1,1,1,1), q_2 \rangle \, q_2 \\ &= (1,1,1,1) - \frac{1}{2} \langle (1,1,1,1), (1,1,0,0) \rangle \, (1,1,0,0) \\ &\quad - \frac{1}{19/2} \langle (1,1,1,1), (-1/2,1/2,3,0) \rangle \, (-1/2,1/2,3,0) \\ &= (1,1,1,1) - (1,1,0,0) - (6/19)(-1/2,1/2,3,0) \\ &= (3/19, -3/19, 1/19, 1) \end{aligned}$$

$$\boxed{q_3 = \frac{Q_3}{||Q_3||} = \frac{(3/19,-3/19,1/19,1)}{||(3/19,-3/19,1/19,1)||} = \frac{(3/19,-3/19,1/19,1)}{\sqrt{20/19}}}$$

We now check for errors:

$$\begin{aligned} ||q_3||^2 &= \frac{(3/19)^2 + (3/19)^2 + (1/19)^2 + 1}{20/19} = 1 \\ q_1^\top q_3 &= \frac{1}{\sqrt{2}\sqrt{20/19}}(3/19 - 3/19) = 0 \\ q_2^\top q_3 &= \frac{1}{\sqrt{19/2}\sqrt{20/19}}((-1/2)(3/19) + (1/2)(-3/19) + (3)(1/19)) = 0 \end{aligned}$$

Thus, $\{q_1, q_2, q_3\}$ is an orthonormal basis for S.

We could also use a computer to crunch the numbers. Enclosed is a MATLAB implementation that uses finite decimal precision to produce the following output, which is consistent with the above calculations:

$q_1 = (0.707106781186547, 0.707106781186547, 0, 0)$

$q_2 = (-0.162221421130762, 0.162221421130763, 0.973328526784575, 0)$

$q_3 = (0.153896752812773, -0.153896752812773, 0.051298917604258, 0.974679434480897)$

Listing 8.1. MATLAB for Gram-Schmidt

```
format long;

v1=[1;1;0;0];
v2=[1;2;3;0];
v3=[1;1;1;1];

V=[v1 v2 v3];
q=zeros(size(V));

for i=1:3
  v=V(:,i);
  val = 0;
  for j=1:i-1
     qj = q(:,j);
     val = val +(v'*qj)*qj;
  end

  Qi=v-val;
  qi=Qi/norm(Qi);
  q(:,i)=qi;
end

q1=q(:,1)
q2=q(:,2)
q3=q(:,3)
```

8.9.3. Example: Gram-Schmidt for polynomial functions

Problem: Let V be the real vector space of square integrable functions $f : [0,1] \to \mathbb{R}$. Consider the standard inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$ for $f, g \in V$. Use the Gram-Schmidt procedure to find an orthonormal basis for the subspace S of all polynomials of degree less than or equal to 2.

Solution: We know $S = Span\{1, t, t^2\}$. We use $v_1 = 1, v_2 = t, v_3 = t^2$. Define

$$\boxed{q_1 = \frac{1}{||1||} = 1}$$

where we use $||1||^2 = \int_0^1 1^2 dt = 1$. Define

$$Q_2 = v_2 - \langle v_2, q_1 \rangle q_1 = t - \langle t, 1 \rangle \cdot 1 = t - 1/2$$

where we use

$$\langle t, 1 \rangle = \int_0^1 t \cdot 1 dt = 1/2$$

Then

$$\boxed{q_2 = \frac{Q_2}{||Q_2||} = \frac{t - 1/2}{\sqrt{\int_0^1 (t - 1/2)^2 dt}} = \frac{t - 1/2}{\sqrt{1/12}}}$$

We can verify correctness of calculations so far:

$$\begin{aligned}
||q_1||^2 &= \int_0^1 1^2 dt = 1 \\
||q_2||^2 &= \int_0^1 \frac{(t - 1/2)^2}{1/12} dt = 1 \\
\langle q_1, q_2 \rangle &= \int_0^1 1 \cdot \frac{t - 1/2}{\sqrt{1/12}} dt = 0
\end{aligned}$$

Define

$$\begin{aligned}
Q_3 &= v_3 - \langle v_3, q_1 \rangle q_1 - \langle v_3, q_2 \rangle q_2 \\
&= t^2 - \left\langle t^2, 1 \right\rangle 1 - \left\langle t^2, \frac{t - 1/2}{\sqrt{1/12}} \right\rangle \frac{t - 1/2}{\sqrt{1/12}} \\
&= t^2 - (1/3) - (t - 1/2) \\
&= t^2 - t + 1/6
\end{aligned}$$

where we use

$$\begin{aligned}
\left\langle t^2, 1 \right\rangle &= \int_0^1 t^2 \cdot 1 dt = 1/3 \\
\left\langle t^2, \frac{t - 1/2}{\sqrt{1/12}} \right\rangle &= \sqrt{12} \int_0^1 t^2 (t - 1/2) dt = \sqrt{1/12}
\end{aligned}$$

So

$$\boxed{q_3 = \frac{Q_3}{\|Q_3\|} = \frac{t^2 - t + 1/6}{\sqrt{\int_0^1 (t^2 - t + 1/6)^2 dt}} = \frac{30}{\sqrt{5}}(t^2 - t + 1/6)}$$

So $\{q_1, q_2, q_3\}$ is an orthonormal basis for S. We can verify correctness:

$$\begin{aligned}
\|q_3\|^2 &= \frac{30^2}{5} \int_0^1 (t^2 - t + 1/6)^2 dt = 1 \\
\langle q_1, q_3 \rangle &= \frac{30}{\sqrt{5}} \int_0^1 1 \cdot (t^2 - t + 1/6) dt = 0 \\
\langle q_2, q_3 \rangle &= \frac{30}{\sqrt{5/12}} \int_0^1 (t - 1/2) \cdot (t^2 - t + 1/6) dt = 0
\end{aligned}$$

8.10. Existence results

8.10.1. Every basis for $\mathbb{R}^n$ can be an orthonormal basis

Let $\{e_1, \ldots, e_n\}$ be the standard basis for $\mathbb{R}^n$, so e_i is the ith column of the $n \times n$ identity matrix. Under the standard inner product for $\mathbb{R}^n$, $\{e_1, \ldots, e_n\}$ is an orthonormal basis:

$$e_i^\top e_j = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases} \tag{8.26}$$

Here is a surprising result.

THEOREM 8.2. *Let $\{x_1, \ldots, x_n\}$ be any basis for $\mathbb{R}^n$. Then, there exists an inner product on $\mathbb{R}^n$ under which $\{x_1, \ldots, x_n\}$ is an orthonormal basis for $\mathbb{R}^n$. Specifically, let X be the real $n \times n$ matrix with columns $\{x_1, \ldots, x_n\}$ and define*

$$A = (X^{-1})^\top X^{-1}$$

Then A is symmetric and positive definite, and $\{x_1, \ldots, x_n\}$ are orthonormal under inner product $\langle x, y \rangle = x^\top A y$ for $x, y \in \mathbb{R}^n$.

PROOF. The square matrix X is invertible because it has linearly independent columns, so the definition of A makes sense. Exercise 8.37 shows that A is symmetric and positive definite. Observe that

$$(X^{-1}X = I) \iff X^{-1} \begin{bmatrix} | & | & \cdots & | \\ x_1 & x_2 & \cdots & x_n \\ | & | & \cdots & | \end{bmatrix} = \begin{bmatrix} | & | & \cdots & | \\ e_1 & e_2 & \cdots & e_n \\ | & | & \cdots & | \end{bmatrix}$$

and so

$$X^{-1}x_i = e_i \quad \forall i \in \{1, ..., n\} \tag{8.27}$$

Fix $i, j \in \{1, ..., n\}$. Then

$$\begin{aligned} x_i^\top A x_j &= x_i^\top (X^{-1})^\top X^{-1} x_j \\ &= (X^{-1}x_i)^\top (X^{-1}x_j) \\ &= e_i^\top e_j \end{aligned}$$

where the final equality holds by (8.27). The result follows by (8.26). □

COROLLARY 8.1. *If $\{x_1, ..., x_k\}$ are vectors in $\mathbb{R}^n$, then $\{x_1, ..., x_k\}$ are linearly independent if and only if there exists an inner product on $\mathbb{R}^n$ under which they are orthonormal.*

PROOF. The reverse direction follows by Lemma 8.7 (orthonormal implies linear independence). For the forward direction, suppose $\{x_1, ..., x_k\}$ are linearly independent. Extend these to a basis for $\mathbb{R}^n$ by adding vectors $\{x_{k+1}, ..., x_n\}$. Then use Theorem 8.2 to create an inner product under which $\{x_1, ..., x_n\}$ are orthonormal. □

8.10.2. Every real vector space has a real inner product

Let V be a real vector space with $\dim(V) = n$ for some $n \in \mathbb{N}$. Let $\{v_1, ..., v_n\}$ be a basis for V. Define the function $T : V \to \mathbb{R}^n$ by mapping each $v \in V$ to the vector $x = (x_1, ..., x_n) \in \mathbb{R}^n$, where $(x_1, ..., x_n)$ are the unique coefficients that satisfy

$$v = \sum_{i=1}^{n} x_i v_i \tag{8.28}$$

It can be shown that T is an invertible linear transformation. Now let A be any real $n \times n$ symmetric and positive definite matrix (such as $A = I$). Exercise 8.38 shows that the following is an inner product on V:

$$\langle v, w \rangle = T(v)^\top A T(w) \quad \forall v, w \in V \tag{8.29}$$

Thus, V has an inner product. (Exercise 8.39 shows that all inner products on V have the form (8.29) for some real $n \times n$ symmetric and positive definite matrix A.)

What about the trivial vector space $V = \{0\}$? Well, that is trivial (define the inner product $\langle 0, 0 \rangle = 0$). What about infinite dimensional spaces? Appendix E uses the concept of *Hamel basis* to show existence of an inner product for any infinite dimensional real vector space. However, this uses a set theory axiom called the *axiom of choice* which

does not tell how to find the inner product. For example, the real vector space of *all functions* $f : [0,1] \to \mathbb{R}$ technically has an inner product, but nobody can compute it since nobody knows what the Hamel basis looks like. It is much nicer to restrict attention to the subspace of *square integrable* functions, for which there is a natural inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$.

The following is a corollary to Theorem 8.2. The corollary treats finite dimensional vector spaces, so its proof does not need a Hamel basis.

COROLLARY 8.2. *If V is a real vector space with $\dim(V) = n$ for some $n \in \mathbb{N}$, then every basis for V is an orthonormal basis under some inner product for V. Further, if $\{x_1, ..., x_k\}$ are vectors in V, then $\{x_1, ..., x_k\}$ are linearly independent if and only if there is an inner product on V under which they are orthonormal.*

PROOF. Let $\{v_1, ..., v_n\}$ be a basis for V. Define the invertible linear transformation $T : V \to \mathbb{R}^n$ that maps each $v \in V$ to the vector $T(v) = (x_1, ..., x_n) \in \mathbb{R}^n$ of unique coefficients that satisfy (8.28). Since $T : V \to \mathbb{R}^n$ is an invertible linear transformation, it follows that $\{T(v_1), ..., T(v_n)\}$ is a basis for $\mathbb{R}^n$. Theorem 8.2 ensures there is a real positive definite matrix A such that $T(v_i)^\top A T(v_j) = 1_{\{i=j\}}$, where $1_{\{i=j\}}$ is the indicator function that is 1 if $i = j$, and 0 else. Define $\langle u, v \rangle = T(u)^\top A T(v)$ for all $u, v \in V$. Then $\langle u, v \rangle$ is a real inner product on V under which $\{v_1, ..., v_n\}$ are orthonormal.

The necessary and sufficient condition for linear independence of vectors $\{x_1, ..., x_k\}$ follows by an argument similar to Corollary 8.1. □

8.11. Perpendicular spaces

Let S be a subspace of a real inner product space V. Define the set $S^\perp$, pronounced "S-perp," as the set of all vectors in V that are orthogonal to all vectors in S

$$S^\perp = \{w \in V : \langle w, s \rangle = 0 \quad \forall s \in S\}$$

The set $S^\perp$ is nonempty because $0 \in S^\perp$. The set $S^\perp$ is called the *orthogonal complement* of subspace S. The next lemma shows that $S^\perp$ is itself a subspace. It is often called the subspace that is *perpendicular* to S.

LEMMA 8.8. *Let S be a subspace of a real inner product space V. Then*

a) If $s \in S$ and $w \in S^\perp$, then $\langle s, w \rangle = 0$.

b) $S^\perp$ is a subspace of V.

c) $S \cap S^\perp = \{0\}$.

d) $S \subseteq (S^\perp)^\perp$.

PROOF. Parts (a)-(c) are in Exercise 8.17. To show part (d), fix $s \in S$. We want to show s is

orthogonal to all vectors in $S^\perp$. Let w be a vector in $S^\perp$. It suffices to show s is orthogonal to w. Since $w \in S^\perp$, we know w is orthogonal to all vectors in S, including s. □

The next lemma provides a simple characterization of $S^\perp$ in the case when S can be described as the span of a collection C of vectors: Instead of establishing $w \in S^\perp$ by showing w is orthogonal to *all* vectors in S, it is enough to show w is orthogonal to all vectors in C.

LEMMA 8.9. *Let S be a subspace of a real inner product space V.*

a) If $S = Span\{v_1, ..., v_n\}$ for some vectors v_i then

$$S^\perp = \{w \in V : \langle w, v_i \rangle = 0 \quad \forall i \in \{1, ..., n\}\}$$

b) If C is a (possibly infinite) collection of vectors in V and $S = Span(C)$, where $Span(C)$ denotes the set of all linear combinations of vectors in C, then $S^\perp = \{w \in V : \langle w, v \rangle = 0 \quad \forall v \in C\}$.

PROOF. The proof of part (a) is in Exercise 8.20. The proof of part (b) is similar. □

The next lemma gives extra results that hold when $\dim(S) < \infty$. It uses the concept of *direct sum* defined in Chapter 6.16.

LEMMA 8.10. *Let S be a* finite dimensional *subspace of a real inner product space V. (The space V can be infinite dimensional.) Then*

a) Every $v \in V$ can be uniquely written $v = s + w$ for some $s \in S$ and $w \in S^\perp$.

b) V is the direct sum $V = S \oplus S^\perp$.

c) $\dim(S) + \dim(S^\perp) = \dim(V)$. [So $(\dim(S^\perp) = \infty) \iff (\dim(V) = \infty)$]

d) $(S^\perp)^\perp = S$.

PROOF. To prove (a), first consider uniqueness: Fix $v \in V$ and suppose

$$v = s_1 + w_1 = s_2 + w_2 \tag{8.30}$$

for some $s_1, s_2 \in S$ and $w_1, w_2 \in S^\perp$. Then

$$s_1 - s_2 = w_2 - w_1 \in S \cap S^\perp = \{0\} \tag{8.31}$$

which uses the fact that S and $S^\perp$ are subspaces so

$$(s_1, s_2 \in S) \implies (s_1 - s_2 \in S)$$
$$(w_1, w_2 \in S^\perp) \implies (w_2 - w_1 \in S^\perp)$$

The final equality in (8.31) is due to the fact that the only vector in both S and $S^\perp$ is 0. So (8.31) implies $s_1 = s_2$ and $w_1 = w_2$. The decomposition in (8.30) is unique.

To complete the proof of (a), fix $v \in V$. We want to show $v = s + w$ for some $s \in S$ and

$w \in S^\perp$. The result is trivial if $S = \{0\}$ (since then $S^\perp = V$ and $v = 0 + v$). Suppose S is nontrivial. Since $\dim(S) < \infty$, the Gram-Schmidt theorem (Theorem 8.1) ensures S has an orthonormal basis $\{q_1, \ldots, q_n\}$, where $n = \dim(S)$. Define

$$s = \sum_{i=1}^{n} \langle v, q_i \rangle \, q_i$$

$$w = v - s$$

The reader can verify that: (i) $v = s + w$; (ii) $s \in S$; (iii) $w \in S^\perp$. [To show $w \in S^\perp$, it is enough to show $\langle w, q_j \rangle = 0$ for all $j \in \{1, \ldots, n\}$.] This completes the proof of part (a).

Part (b) is a direct consequence of (a). The proof of part (c) uses (a) with Lemma 6.7. To show part (d), observe that Lemma 8.8(d) ensures $S \subseteq (S^\perp)^\perp$. It suffices to show the reverse inclusion $(S^\perp)^\perp \subseteq S$. Fix $v \in (S^\perp)^\perp$. We want to show $v \in S$. Part (b) implies $v = s + w$ for some $s \in S$ and $w \in S^\perp$. It suffices to show $w = 0$. Since $w \in S^\perp$ and $v \in (S^\perp)^\perp$ we know

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

Substituting $v = s + w$ gives

$$\langle w, s + w \rangle = 0$$

Thus

$$0 = \langle w, s + w \rangle = \langle w, s \rangle + \langle w, w \rangle = 0 + ||w||^2$$

where the final equality uses $\langle w, s \rangle = 0$ since $s \in S$ and $w \in S^\perp$. Thus $||w||^2 = 0$, so $w = 0$. □

8.11.1. Examples of perpendicular spaces

1. (Finding $S^\perp$ when $\dim(S) = 1$) Consider $\mathbb{R}^2$ with the standard inner product

$$\langle x, y \rangle = x^\top y = x_1 y_1 + x_2 y_2 \quad \forall x, y \in \mathbb{R}^2$$

Define $S = Span\{(3, 7)\}$. What is $S^\perp$? Lemma 8.9(a) gives

$$S^\perp = \left\{ w \in \mathbb{R}^2 : w^\top \begin{bmatrix} 3 \\ 7 \end{bmatrix} = 0 \right\} = \{(w_1, w_2) \in \mathbb{R}^2 : 3w_1 + 7w_2 = 0\}$$

Observe that $(-7, 3) \in S^\perp$. So $S^\perp$ is a nontrivial subspace, but is not *all* of $\mathbb{R}^2$, so

$$\left(\{0\} \subsetneq S^\perp \subsetneq \mathbb{R}^2 \right) \implies \left(0 < \dim(S^\perp) < 2 \right)$$

It follows that $\dim(S^\perp) = 1$ and $S^\perp = Span\{(-7, 3)\}$. Of course, we could have concluded $\dim(S^\perp) = 1$ directly from Lemma 8.10(c).

2. (Finding $Col(A)^\perp$) Consider the real vector space $\mathbb{R}^4$ with the standard inner product $x^\top y$ for all $x, y \in \mathbb{R}^4$. Define

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

Since $Col(A) = Span\{(2,3,0,4),(0,1,4,1)\}$, by Lemma 8.9(a) we know

$$Col(A)^\perp = \left\{ w \in \mathbb{R}^4 : w^\top \begin{bmatrix} 2 \\ 3 \\ 0 \\ 4 \end{bmatrix} = 0 \text{ and } w^\top \begin{bmatrix} 0 \\ 1 \\ 4 \\ 1 \end{bmatrix} = 0 \right\}$$

Thus,

$$(w \in Col(A)^\perp) \iff \left(\underbrace{\begin{bmatrix} 2 & 3 & 0 & 4 \\ 0 & 1 & 4 & 1 \end{bmatrix}}_{A^\top} w = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right) \iff (w \in Null(A^\top))$$

So $Col(A)^\perp = Null(A^\top)$. Lemma 8.10(c) ensures

$$\underbrace{\dim(Col(A))}_{2} + \dim(Col(A)^\perp) = 4$$

which means $\dim(Col(A)^\perp) = 2$, which can be verified by finding $\dim Null(A^\top)$.

3. ($S^\perp$ with functions) Let V be the real vector space of square integrable functions $f : [0,1] \to \mathbb{R}$. Use the standard inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$ for $f, g \in V$. Define

$$S = Span\{t, e^t\}$$

By Lemma 8.9(a) we have

$$S^\perp = \left\{ f \in V : \int_0^1 f(t)t\,dt = 0 \text{ and } \int_0^1 f(t)e^t dt = 0 \right\}$$

It is not surprising that $\dim(S^\perp) = \infty$, which is guaranteed by Lemma 8.10(c). Exercise 8.43 manufactures a wide collection of example functions $f \in S^\perp$.

4. (Odd and even functions) Let V be the real vector space of continuous functions $f : [-1, 1] \to \mathbb{R}$. Consider the standard inner product $\langle f, g \rangle = \int_{-1}^{1} f(t)g(t)dt$ for $f, g \in V$. [We have restricted attention to the *continuous* functions because it removes the need for *equivalence classes*: Any two functions that are continuous and equivalent must be the same function.] Now let S_{even} and S_{odd} be the subspaces of even and odd functions:

$$S_{even} = \{f \in V : f(-t) = f(t) \quad \forall t \in [-1, 1]\}$$
$$S_{odd} = \{f \in V : f(-t) = -f(t) \quad \forall t \in [-1, 1]\}$$

The functions $\{1, t^2, t^4, t^6, \ldots\}$ are all in S_{even}, so $\dim(S_{even}) = \infty$. Since Lemma 8.10(d) requires $\dim(S) < \infty$, we cannot immediately conclude that $(S_{even}^{\perp})^{\perp} = S_{even}$. However, it turns out to be true in this case. Indeed, it can be shown that

$$S_{even}^{\perp} = S_{odd} \quad , \quad S_{odd}^{\perp} = S_{even}$$

Further, it holds that $V = S_{even} \oplus S_{odd}$. The steps to show these things are as follows: Exercise 9.42 shows that every $f \in V$ can be uniquely written $f = f_{even} + f_{odd}$, where $f_{even} \in S_{even}$ and $f_{odd} \in S_{odd}$ are defined as the *even part* and *odd part* of f:

$$f_{even}(t) = \frac{f(t) + f(-t)}{2} \quad \forall t \in [-1, 1]$$
$$f_{odd}(t) = \frac{f(t) - f(-t)}{2} \quad \forall t \in [-1, 1]$$

Also, Exercise 9.43 shows that if $f \in S_{even}$ and $g \in S_{odd}$ then $\langle f, g \rangle = 0$. We now use these facts to prove $S_{even} = S_{odd}^{\perp}$.

Claim 1: $S_{even} \subseteq S_{odd}^{\perp}$. Fix $f \in S_{even}$. We want to show f is orthogonal to all functions in S_{odd}. We already know this from the above discussion, so we are done.
Claim 2: $S_{even} \supseteq S_{odd}^{\perp}$. Let f be any function in $S_{odd}^{\perp}$. We want to show $f \in S_{even}$. From the above discussion, we know $f = f_{even} + f_{odd}$ for some $f_{even} \in S_{even}$ and $f_{odd} \in S_{odd}$. But $f \in S_{odd}^{\perp}$, so f is orthogonal to all functions in S_{odd}, so f is orthogonal to f_{odd}:

$$\langle f, f_{odd} \rangle = 0 \implies \langle f_{even} + f_{odd}, f_{odd} \rangle = 0 \implies \underbrace{\langle f_{even}, f_{odd} \rangle}_{0} + \underbrace{\langle f_{odd}, f_{odd} \rangle}_{\|f_{odd}\|^2} = 0$$

So $\|f_{odd}\|^2 = 0$, hence $f_{odd} = 0$ and $f = f_{even} \in S_{even}$. We are done.

8.11.2. A case when $S \subsetneq (S^\perp)^\perp$

It is possible to have $S \subsetneq (S^\perp)^\perp$. This can only happen when $\dim(S) = \infty$. An example is when V is the real inner product space of square integrable functions from $[0,1]$ to $\mathbb{R}$:

$$V = \left\{ f : [0,1] \to \mathbb{R} : \int_0^1 f(t)^2 dt < \infty \right\}$$

Let $S \subseteq V$ be the (infinite dimensional) subspace of polynomial functions. Consider $e^t \in V$, where e^t represents the exponential function defined over $t \in [0,1]$. We know $e^t \notin S$ since e^t is not polynomial. We now show $e^t \in (S^\perp)^\perp$. To show e^t is orthogonal to everything in $S^\perp$, fix $z \in S^\perp$. It suffices to show $\langle e^t, z \rangle = 0$. Define polynomials p_n for $n \in \mathbb{N}$ by

$$p_n(t) = \sum_{k=0}^{n} \frac{t^k}{k!} \quad \forall t \in [0,1]$$

It can be shown that

$$\lim_{n \to \infty} \left\| e^t - p_n \right\| = 0 \tag{8.32}$$

Fix $n \in \mathbb{N}$. Since $p_n \in S$ and $z \in S^\perp$ we know $\langle p_n, z \rangle = 0$. Then

$$\begin{aligned} \langle e^t, z \rangle &= \langle e^t - p_n + p_n, z \rangle \\ &= \langle e^t - p_n, z \rangle + \langle p_n, z \rangle \\ &\overset{(a)}{=} \langle e^t - p_n, z \rangle + 0 \\ &\leq \left\| e^t - p_n \right\| \cdot \|z\| \end{aligned}$$

where (a) uses $\langle p_n, z \rangle = 0$; the final inequality holds by Cauchy-Schwarz. This holds for all $n \in \mathbb{N}$. Taking $n \to \infty$ and using (8.32) gives

$$\langle e^t, z \rangle \leq 0$$

A similar argument that computes $\langle e^t, -z \rangle$ can show $\langle e^t, z \rangle \geq 0$, so we must have $\langle e^t, z \rangle = 0$. See Exercise 8.29 for another example.

The example situation of the previous paragraph is possible because the function e^t *cannot* be written in the form $e^t = s + w$ for some $s \in S$ and $w \in S^\perp$. Thus, that same example shows how part (a) of Lemma 8.10 can fail if S is infinite dimensional. If S is infinite dimensional but has the property that every $v \in V$ can be written $v = s + w$ for some $s \in S$ and $w \in S^\perp$, then the proof of Lemma 8.10(d) can be repeated to show $(S^\perp)^\perp = S$.

8.11.3. Perpendicular spaces for matrices

Consider the standard inner product $x^\top y$ in $\mathbb{R}^m$ and $\mathbb{R}^n$.

LEMMA 8.11. *If A is a $m \times n$ matrix with real entries, then*

$$Col(A)^\perp = Null(A^\top) \tag{8.33}$$

$$Null(A)^\perp = Col(A^\top) \tag{8.34}$$

Further, $Col(A) = Null(A^\top)^\perp$ and $Null(A) = Col(A^\top)^\perp$.

We give two different proofs of Lemma 8.11: One that proves (8.33); another that proves (8.34). Note that proving *either* (8.33) *or* (8.34) immediately yields the remaining three equalities of the lemma. This is because the null or column space of a matrix is finite dimensional, so we can use the general fact $(S^\perp)^\perp = S$ for subspaces S that satisfy $\dim(S) < \infty$. For example, suppose we can prove that (8.33) is true. Taking the perp of both sides of (8.33) yields

$$Col(A) = Null(A^\top)^\perp$$

and since this equality holds for any matrix A, it also holds for matrix $A^\top$, so

$$Col(A^\top) = Null(A)^\perp$$

and taking the perp of both sides gives

$$Col(A^\top)^\perp = Null(A)$$

which establishes all four equalities of the lemma, assuming (8.33) holds. The same could be done if we assume (8.34) holds.

Proof 1 of Lemma 8.11: To prove (8.33), let $\{c_1, ..., c_n\}$ be the columns of A. Then $Col(A) = Span\{c_1, ..., c_n\}$ and Lemma 8.9(a) implies

$$Col(A)^\perp = \{w : w^\top c_i = 0 \quad \forall i \in \{1, ..., n\}\}$$

Thus, $w \in Col(A)^\perp$ if and only if w satisfies the matrix equation

$$\underbrace{\begin{bmatrix} - & c_1^\top & - \\ - & c_2^\top & - \\ \vdots & \vdots & \vdots \\ - & c_n^\top & - \end{bmatrix}}_{A^\top} w = 0$$

In particular, $w \in Col(A)^\perp$ if and only if $w \in Null(A^\top)$, that is,

$$Col(A)^\perp = Null(A^\top) \tag{8.35}$$

This proves (8.33), from which the remaining three equalities of the lemma follow. □

Proof 2 of Lemma 8.11: To prove (8.34), we first show $Col(A^\top) \subseteq Null(A)^\perp$. Fix $w \in Col(A^\top)$. We want to show w is orthogonal to all vectors in $Null(A)$. Let z be a vector in $Null(A)$. It suffices to show $w^\top z = 0$. Since $w \in Col(A^\top)$ we know

$$w = A^\top y$$

for some vector y. Taking transposes gives

$$w^\top = y^\top A$$

Multiplying both sides by z gives

$$w^\top z = y^\top A z = 0$$

where we have used $Az = 0$ (since $z \in Null(A)$).

To complete the proof of (8.34), it remains to show there are no other points in $Null(A)^\perp$ that are not already in $Col(A^\top)$. We use a dimensionality argument: Since the two subspaces satisfy $Col(A^\top) \subseteq Null(A)^\perp$, if they have the same dimension they must be the same subspace (recall Lemma 6.4). Lemma 8.10(c) gives

$$\dim(Null(A)) + \dim(Null(A)^\perp) = n$$

Thus

$$\begin{aligned} \dim(Null(A)^\perp) &= n - \dim(Null(A)) \\ &= \operatorname{rank}(A) \quad \text{[rank-nullity theorem]} \\ &= \operatorname{rank}(A^\top) \quad \text{[Lemma 5.4]} \\ &= \dim Col(A^\top) \end{aligned}$$

This establishes (8.34), from which the remaining three equalities of the lemma follow. □

8.12. Gram matrix

Let $\{v_1, \ldots, v_n\}$ be vectors in a real inner product space V.

DEFINITION 8.3. *The* Gram matrix *associated with the vectors* $\{v_1, \ldots, v_n\}$ *is defined as the*

$n \times n$ matrix of pairwise inner products:

$$G = \left(\left\langle v_i, v_j \right\rangle\right)$$

so the (i, j) entry of G is $g_{ij} = \left\langle v_i, v_j \right\rangle$.

Inner products between linear combinations of vectors in $\{v_1, ..., v_n\}$ can be obtained easily from the Gram matrix, as shown in the following lemma.

LEMMA 8.12. *Let G be the Gram matrix for $\{v_1, ..., v_n\}$. Then*

a) For any $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$ and $b = (b_1, ..., b_n) \in \mathbb{R}^n$ we have

$$\left\langle \sum_{i=1}^{n} a_i v_i, \sum_{j=1}^{n} b_j v_j \right\rangle = a^\top G b$$

b) For any $a = (a_1, ..., a_n) \in \mathbb{R}^n$ we have

$$\left\| \sum_{i=1}^{n} a_i v_i \right\|^2 = a^\top G a$$

c) The Gram matrix G is symmetric and positive semidefinite.

PROOF. Let $G = (g_{ij})$. To prove (a), from (8.1) we have

$$\left\langle \sum_{i=1}^{n} a_i v_i, \sum_{j=1}^{n} b_j v_j \right\rangle = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i b_j \underbrace{\left\langle v_i, v_j \right\rangle}_{g_{ij}} = a^\top G b$$

where the final equality uses Lemma 7.2. This proves (a). Part (b) follows immediately from (a). Part (c) follows easily from part (b). □

LEMMA 8.13. *Let G be the Gram matrix for $\{v_1, ..., v_n\}$. Then*

$$(G \textit{ is invertible}) \iff (\{v_1, ..., v_n\} \textit{ is linearly independent})$$

PROOF. First suppose $\{v_1, ..., v_n\}$ are linearly independent. To show G is invertible, it suffices to show $Null(G) = \{0\}$. Fix $a = (a_1, ..., a_n) \in \mathbb{R}^n$ and suppose $Ga = 0$. We want to show $a = 0$. Since $Ga = 0$ we know $a^\top Ga = 0$ and the previous lemma implies

$$0 = a^\top G a = \left\| \sum_{i=1}^{n} a_i v_i \right\|^2$$

Since $\left\| \sum_{i=1}^{n} a_i v_i \right\|^2 = 0$, we know

$$\sum_{i=1}^{n} a_i v_i = 0$$

and so $a_i = 0$ for all i since $\{v_1, ..., v_n\}$ are linearly independent. This completes the proof

that G is invertible.

Now suppose G is invertible. To show $\{v_1, \ldots, v_n\}$ are linearly independent, suppose

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

for some scalars $a_1, \ldots, a_n$. We want to show $a_j = 0$ for all j. Fix $i \in \{1, \ldots, n\}$. Taking inner products of the above with v_i gives

$$\left\langle v_i, \sum_{j=1}^{n} a_j v_j \right\rangle = 0$$

and so

$$\sum_{j=1}^{n} a_j \left\langle v_i, v_j \right\rangle = 0 \tag{8.36}$$

Let $a = (a_1, \ldots, a_n)$. By rules of matrix multiplication we have for each $i \in \{1, \ldots, n\}$

$$(i\text{th row of } Ga) = \sum_{j=1}^{n} a_j g_{ij} = \sum_{j=1}^{n} a_j \left\langle v_i, v_j \right\rangle = 0$$

where the final equality holds by (8.36). So every row of Ga is zero. That is, $Ga = 0$, and multiplying both sides by G^{-1} proves that $a = 0$. □

8.12.1. Discussion

Lemma 8.13 demonstrates the power of an inner product. The $n \times n$ Gram matrix for vectors $\{v_1, \ldots, v_n\}$ is computed by only considering *pairwise relationships* between the vectors. It is remarkable that the pairwise inner product information for $\{v_1, \ldots, v_n\}$ is enough to establish whether or not $\{v_1, \ldots, v_n\}$ is linearly independent.

This is particularly surprising in view of the fact that linear independence of $\{v_1, \ldots, v_n\}$ is a property of the *entire collection* of vectors: Pairwise linear independence is not enough to prove the entire collection is linearly independent. For example, the following five vectors are all *pairwise linearly independent* in $\mathbb{R}^2$ because all vectors are nonzero and no vector is a scalar multiple of any other:

$$\left\{ \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 4 \end{bmatrix} \right\} \tag{8.37}$$

Of course, the entire collection of five vectors cannot be linearly independent because we cannot have a list of five linearly independent vectors in the 2-dimensional space $\mathbb{R}^2$.

8.12.2. Two important cases

- In the special case when vectors $\{v_1, ..., v_n\}$ are orthonormal, the Gram matrix is the $n \times n$ identity matrix. That is, $G = I$.
- Suppose $\{v_1, ..., v_n\}$ are vectors in $\mathbb{R}^m$. Let A be the $m \times n$ matrix with n columns given by $\{v_1, ..., v_n\}$. Assuming the standard inner product $x^\top y$ is used for vectors in $\mathbb{R}^m$, the Gram matrix for $\{v_1, ..., v_n\}$ is the $n \times n$ matrix $A^\top A$. That is, $G = A^\top A$. For example, the 5×5 Gram matrix for the five vectors in $\mathbb{R}^2$ from (8.37) is

$$G = \underbrace{\begin{bmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \end{bmatrix}}_{A^\top} \underbrace{\begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 2 & 3 & 4 \end{bmatrix}}_{A} = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 3 & 5 & 7 & 9 \\ 1 & 4 & 7 & 10 & 13 \\ 1 & 5 & 9 & 13 & 17 \end{bmatrix}$$

  and, of course, this matrix G is not invertible.

8.13. Exercises

EXERCISE 8.1. *(Useful inequalities) Show that for all $x, y \in \mathbb{R}$ and $\theta > 0$:*

a) $(x + y)^2 \leq 2x^2 + 2y^2$

b) $|xy| \leq \frac{1}{2}x^2 + \frac{1}{2}y^2$

c) $|xy| \leq \frac{\theta}{2}x^2 + \frac{1}{2\theta}y^2$

d) $(x + y)^2 \leq (1 + \theta)x^2 + \left(1 + \frac{1}{\theta}\right) y^2$

e) Use one or more of the above to show that if V is the set of functions $f : [0, 1] \to \mathbb{R}$ such that $\int_0^1 f(t)^2 dt < \infty$, then $f, g \in V$ implies $f + g \in V$ and $\int_0^1 f(t)g(t)dt \in \mathbb{R}$.

EXERCISE 8.2. *(Orthogonal to everything) Let V be a real inner product space. Suppose $z \in V$ and z is orthogonal to all vectors in V, so that $\langle z, x\rangle = 0$ for all $x \in V$. What can be said about z?*

EXERCISE 8.3. *(Example of orthogonal) Using the standard inner product in $\mathbb{R}^3$, find a vector $w \in \mathbb{R}^3$ that is orthogonal to both $(1, 1, 2)$ and $(-1, 0, 7)$ and that satisfies $||w|| = 1$. Is your vector w orthogonal to everything in $Span\{(1, 1, 2), (-1, 0, 7)\}$?*

EXERCISE 8.4. *(Orthogonal functions) Let V be the real inner product space of square integrable functions $f : [0, 1] \to \mathbb{R}$ with the standard inner product $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$. Find a real number $a \in \mathbb{R}$ such that t and $e^t - at$ are orthogonal.*

EXERCISE 8.5. *(The inner product $x^\top Ay$ on $\mathbb{R}^n$) Fix n as a positive integer. Let $A \sim n \times n$ be symmetric and positive definite. Define the function $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ by $f(x, y) = x^\top Ay$. This is*

the same function as in Section 7.1, but in the special case when A is square and has real entries.

a) Prove that $f(x, y)$ is an inner product on $\mathbb{R}^n$ by showing it satisfies the nonnegativity, symmetry, and linearity properties.

b) Give an example 2×2 matrix A that is not the identity matrix, and is not the $[2\ -1; -1\ 2]$ matrix used as the example in Section 7.2, but that is symmetric and positive definite. Numerically compute $f(x, y)$ for this example for four different choices of x, y, where at least one of your choices has $x = y \neq 0$.

EXERCISE 8.6. *(Cauchy-Schwarz $\implies$ Triangle inequality) Let $\langle x, y \rangle$ be an inner product on a vector space V. Define $||x|| = \sqrt{\langle x, x \rangle}$ for all $x \in V$. By computing $||x + y||^2$, use the Cauchy-Schwarz inequality to prove the triangle inequality:*

$$||x + y|| \leq ||x|| + ||y|| \quad \forall x, y \in V$$

EXERCISE 8.7. *(Strengthened Cauchy-Schwarz for random variables) Consider a particular probability space and let V be the set of all random variables on this space that have finite second moments.*

a) Show that $X \in V \implies |X| \in V$.

b) Use part (a) to strengthen the inequality (8.14) *to $\mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}\left[X^2\right]\mathbb{E}\left[Y^2\right]} \quad \forall X, Y \in V$.*

EXERCISE 8.8. *(Equality in Cauchy-Schwarz) Let V be a real inner product space. Let x, y be vectors in V. Show the Cauchy-Schwarz inequality $|\langle x, y \rangle| \leq ||x|| \cdot ||y||$ holds with equality if and only if either $y = 0$ or $x = ay$ for some $a \in \mathbb{R}$.* Hint: One direction is easy. For the other direction, you may need to look back on the proof of the Cauchy-Schwarz inequality.

EXERCISE 8.9. *(Different inner products) If two vectors in a real vector space are orthogonal under one inner product, must they also be orthogonal in another inner product? Prove true, or give a counter-example.*

EXERCISE 8.10. *(Orthogonal subspaces) If M and N are two subspaces of a real inner product space, prove:*

$$M \subseteq N \implies N^{\perp} \subseteq M^{\perp}$$

EXERCISE 8.11. *(Basis for orthogonal complement) Find a basis for the orthogonal complement of M, where*

$$M = Span\{(1, 2, 0, 3), (2, 4, 1, 6)\}$$

EXERCISE 8.12. *Find a solution x to the following system of equations that satisfies $||x|| = \sqrt{300}$:*

$$\begin{bmatrix} 3 & 3 & 2 \\ 2 & -5 & 2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

where we use field $\mathbb{R}$ and the standard inner product norm on $\mathbb{R}^3$.

EXERCISE 8.13. *Find a basis for $Col(A)^\perp$ when A is the following matrix with real entries:*

$$A = \begin{bmatrix} 3 & 1 \\ 0 & -1 \\ 2 & 2 \end{bmatrix}$$

EXERCISE 8.14. *(Multiple choice) Suppose we know $y \in Span\{v_1, v_2, v_3\}$. What do we know about y?*

a) Nothing.

b) $y = \sum_{i=1}^{3} a_i v_i$ for some scalars a_i.

c) $\{v_1, v_2, v_3\}$ are linearly independent.

d) $\{y, v_1, v_2, v_3\}$ span the vector space.

EXERCISE 8.15. *(Multiple choice) Suppose we have vectors $\{x_1, ..., x_k\}$ in a vector space U and $\{y_1, ..., y_k\}$ in a vector space V. We know $\{y_1, ..., y_k\}$ are linearly independent. We also know $T(x_i) = y_i$ for all $i \in \{1, ..., k\}$ for some particular linear transformation $T : U \to V$. We want to prove $\{x_1, ..., x_k\}$ are linearly independent. What is the first line of the proof?*

a) Suppose $\sum_{i=1}^{k} a_i x_i = 0$ for some scalars a_i.

b) Suppose $\sum_{i=1}^{k} a_i y_i = 0$ for some scalars a_i.

c) Suppose $T(x_i) = 0$ for all i.

d) Suppose $Null(T)$ is not trivial.

EXERCISE 8.16. *(Multiple choice 3) Suppose $\{x_1, ..., x_k\}$ are nonzero vectors that are pairwise orthogonal. We know*

a) $\{x_1, ..., x_k\}$ are linearly independent.

b) $\left\langle x_i, x_j \right\rangle = 0$ for all $i, j \in \{1, ..., k\}$ such that $i \neq j$.

c) $\left\| \sum_{i=1}^{k} x_i \right\|^2 = \sum_{i=1}^{k} \|x_i\|^2$

d) All of the above

EXERCISE 8.17. *($S^\perp$ is a subspace) Let V be a real inner product space. Let S be a subspace of V.*

a) Show that $S^\perp$ is a subspace of V.

b) Show that if $s \in S$ and $w \in S^\perp$, then $\langle s, w \rangle = 0$. Explain why $S \cap S^\perp = \{0\}$.

EXERCISE 8.18. *(Finding $S^\perp$) Let V be the real inner product space of polynomial functions $f : [0, 1] \to \mathbb{R}$, using the standard inner product $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$ for all $f, g \in V$. Let S be the subspace of all polynomials of degree 2 or less.*

a) Find $S^\perp$.

b) Find a particular nonzero polynomial in $S^\perp$.

EXERCISE 8.19. *(Again finding $S^\perp$) Let V be the real vector space of square integrable functions*

$f : [0, 2\pi] \to \mathbb{R}$, using the standard inner product $\langle f, g\rangle = \int_0^{2\pi} f(t)g(t)dt$ for all $f, g \in V$. Let $S = Span\{\cos(t), t\}$.

a) Find $S^\perp$.

b) Find a particular nonzero function in $S^\perp$.

c) Is $\sin(t) \in S^\perp$?

EXERCISE 8.20. *Let $\{v_1, ..., v_k\}$ be vectors in a real inner product space V. Fix $w \in V$. Show that w is orthogonal to all vectors in $Span\{v_1, ..., v_k\}$ if and only if $\langle w, v_i\rangle = 0$ for all $i \in \{1, ..., k\}$.*

EXERCISE 8.21. *Let V be a real inner product space. Assume $\{v_1, ..., v_k\}$ are orthonormal vectors in V. Assume $k \geq 3$.*

a) Find $\left\|\sum_{i=1}^k v_i\right\|^2$.

b) Find $\left\|\sum_{i=1}^k iv_i\right\|^2$.

c) Find $\|8v_1 - v_2 + 4v_3\|^2$.

EXERCISE 8.22. *(Orthonormal matrix) Suppose A is a real $m \times n$ matrix with orthonormal columns.*

a) Find the size of $A^\top A$. Show that $A^\top A = I$.

b) Show that $\|Ax\|^2 = \|x\|^2$ for all $x \in \mathbb{R}^n$, where we use the standard norm in $\mathbb{R}^n$ and $\mathbb{R}^m$.

c) Fix $\theta \in \mathbb{R}$. Show that the following rotation matrix *has orthonormal columns:*

$$A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$$

Using $\theta = \pi/4$, plot x and Ax for three choices of $x \in \mathbb{R}^2$ such that $\|x\| = 1$.

EXERCISE 8.23. *(Rank 1 matrix) Let m, n be positive integers. Consider field $\mathbb{R}$. Fix $u \in \mathbb{R}^m$ and $v \in \mathbb{R}^n$. Assume $u \neq 0$ and $v \neq 0$. Define the matrix $A = uv^\top$.*

a) What is the size of A?

b) Prove that $rank(A) = 1$.

EXERCISE 8.24. *Let A be a real 10×20 matrix and suppose that*

$$A = 5u_1v_1^\top + 4u_2v_2^\top$$

where $\{u_1, u_2\}$ are orthonormal vectors in $\mathbb{R}^{10}$, $\{v_1, v_2\}$ are orthonormal vectors in $\mathbb{R}^{20}$.

a) Compute Av_1 and Av_2.

b) Compute $A^\top A$. Compute $A^\top Av_1$ and $A^\top Av_2$.

c) Compute $rank(A)$.

EXERCISE 8.25. *(Singular values)[6] Fix positive integers m and n. Let $\{v_1, ..., v_n\}$ be an or-*

[6] Every real and nonzero $m \times n$ matrix A has such a *singular value decomposition* $A = \sum_{i=1}^r \sigma_i u_i v_i^\top$, where

thonormal basis for $\mathbb{R}^n$. Let r be an integer such that $1 \leq r \leq \min\{m, n\}$. Let $\{u_1, ..., u_r\}$ be orthonormal vectors in $\mathbb{R}^m$. Fix σ_i as real numbers that satisfy

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

Define the $m \times n$ matrix $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$.

a) Compute Av_k for all $k \in \{1, ..., n\}$.

b) Find a basis for Null(A).

c) Find rank(A).

d) For real numbers $\theta_1, ..., \theta_n$, compute $\left\|A(\sum_{k=1}^{n} \theta_k v_k)\right\|^2$.

e) Show that $||Ax|| \leq \sigma_1 ||x||$ for all $x \in \mathbb{R}^n$. Hint: $x \in \mathbb{R}^n \implies x \in Span\{v_1, ..., v_n\}$.

f) Suppose $x \in Span\{v_1, ..., v_r\}$. Show that $||Ax|| \geq \sigma_r ||x||$.

EXERCISE 8.26. *(Gram-Schmidt) Use the Gram-Schmidt procedure to turn vectors $\{v_1, v_2, v_3\}$ into orthonormal vectors $\{q_1, q_2, q_3\}$ with the same span. Use the standard inner product on $\mathbb{R}^4$.*

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

EXERCISE 8.27. *(Gram-Schmidt for functions) Let V be the real inner product space of square integrable functions $f : [0, 2\pi] \to \mathbb{R}$. Consider inner product $\langle f, g \rangle = \int_0^{2\pi} f(t)g(t)dt$ for all $f, g \in V$. Carry out the Gram-Schmidt procedure for:*

a) $\{1, e^t\}$

b) $\{\cos(t), \sin(t), t\}$

c) $\{1, t^2, t^4\}$

EXERCISE 8.28. *This problem plays with assumptions in Lemmas 8.5 and 5.4. Define*

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

a) Treat entries of A in the real field. Find $A^\top A$ and $AA^\top$ and verify that

$$rank(A) = rank(A^\top) = rank(A^\top A) = rank(AA^\top)$$

b) Treat entries of A in field $\mathbb{F}_2 = \{0, 1\}$. Find $A^\top A$ and $AA^\top$. Which results of part (a) fail?

$r = \text{rank}(A)$, $\{v_1, ..., v_r\}$ are orthonormal in $\mathbb{R}^n$, $\{u_1, ..., u_r\}$ are orthonormal in $\mathbb{R}^m$. The positive real numbers $\sigma_1, ..., \sigma_r$ are called the *singular values* of matrix A.

EXERCISE 8.29. *($S \neq (S^\perp)^\perp$) Define V as the set of all infinite sequences of real numbers that are square summable:*

$$V = \left\{ (x_1, x_2, x_3, \ldots) \in \mathbb{R}^{\mathbb{N}} : \sum_{i=1}^{\infty} x_i^2 < \infty \right\}$$

Define the real inner product $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$ for all $x, y \in V$. Define

$$S = \{(x_1, x_2, x_3, \ldots) \in V : x_i = 0 \text{ for all but finitely many } i\}$$

a) Show that S is a subspace of V.

b) Show that $S^\perp = \{0\}$, where $0 = (0, 0, 0, \ldots)$ is the zero vector in V. Conclude that $(S^\perp)^\perp = V$ and $S + S^\perp \subsetneq V$.

c) Find a vector w such that $w \in (S^\perp)^\perp$ but $w \notin S$.

d) For your vector w from part (c), show there is no projection of w onto S. That is, for any vector in S, there is another vector in S that is closer to w.

EXERCISE 8.30. *(Finite dimensional subspaces of a real inner product space are closed, and complete) Let V be a real inner product space. Let $\|\cdot\|$ denote the inner product norm. Let S be a finite dimensional subspace of V with orthonormal basis $\{v_1, \ldots, v_n\}$ for some $n \in \mathbb{N}$.*

a) Let $x \in V$ be a limit point of S, so there is a sequence of points $\{s[k]\}_{k=1}^{\infty}$ in S such that $\lim_{k\to\infty} \|s[k] - x\| = 0$. Suppose $x \notin S$. Reach a contradiction by projecting x onto S. Thus, S is a closed subset of V.

b) Suppose $y, z \in S$. Let $y = \sum_{i=1}^{n} a_i v_i$ and $z = \sum_{i=1}^{n} b_i v_i$ be the unique way of writing y, z as linear combinations of $\{v_1, \ldots, v_n\}$. Define $a, b \in \mathbb{R}^n$ by $a = (a_1, \ldots, a_n)$ and $b = (b_1, \ldots, b_n)$. Compute the standard Euclidean distance between a and b in terms of $\|y - z\|$.

c) Use the fact that every Cauchy sequence in $\mathbb{R}^n$ converges to a point in $\mathbb{R}^n$ to show that every Cauchy sequence in S converges to a point in S. Thus, S is complete. *[A sequence $\{x[k]\}_{k=1}^{\infty}$ is a Cauchy sequence if for all $\epsilon > 0$ there is an integer k_ϵ such that the distance between $x[k]$ and $x[r]$ is less than or equal to ϵ whenever $k, r \in \mathbb{N}$ and $k \geq k_\epsilon, r \geq k_\epsilon$.] [Note: For subsets of V, completeness implies closedness, so part (c) is also an alternative proof of closedness.]*

EXERCISE 8.31. *Suppose V is a real inner product space.*

a) Show that $\langle ax, by \rangle = ab \langle x, y \rangle$ for all $a, b \in \mathbb{R}$, $x, y \in V$.

b) Fix $n \in \mathbb{N}$. Use induction to show that $\left\langle x, \sum_{j=1}^{n} y_j \right\rangle = \sum_{j=1}^{n} \langle x, y_j \rangle$ for all $x, y_1, \ldots, y_n \in V$.

c) Fix $m \in \mathbb{N}$. Use part (b) to show $\left\langle \sum_{i=1}^{m} x_i, y \right\rangle = \sum_{i=1}^{m} \langle x_i, y \rangle$ for all $x_1, \ldots, x_m, y \in V$.

d) Fix $m, n \in \mathbb{N}$. Use (b) and (c) to show

$$\left\langle \sum_{i=1}^{m} x_i, \sum_{j=1}^{n} y_j \right\rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} \langle x_i, y_j \rangle \quad \forall x_1, \ldots, x_m, y_1, \ldots, y_n \in V$$

e) Combine parts (d) and (a) to prove (8.1).

EXERCISE 8.32. *(Comparing $||x||_1$, $||x||_2$, $||x||_\infty$) For $\mathbb{R}^n$, the L_1, L_2, and L_∞ norms are defined:*

$$||x||_1 = \sum_{i=1}^{n} |x_i| \quad \forall x \in \mathbb{R}^n$$
$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \quad \forall x \in \mathbb{R}^n$$
$$||x||_\infty = \max\{|x_1|, \ldots, |x_n|\} \quad \forall x \in \mathbb{R}^n$$

[Note that L_1 and L_∞ are not inner product norms.]

a) Use $\sum_{i=1}^{n} |x_i| = \sum_{i=1}^{n} |x_i| \cdot 1$ with the Cauchy-Schwarz inequality for $\mathbb{R}^n$ (see (8.5)) to prove

$$||x||_2 \leq ||x||_1 \leq \sqrt{n}\, ||x||_2 \quad \forall x \in \mathbb{R}^n$$

b) Show that $||x||_\infty \leq ||x||_2 \leq \sqrt{n}\, ||x||_\infty$ for all $x \in \mathbb{R}^n$.

c) Give nontrivial examples where the upper bounds in (a) and (b) hold with equality.

EXERCISE 8.33. *(Triangle inequality) Let V be a real vector space. Let $||\cdot||$ denote a norm on V, not necessarily an inner product norm. [A general norm is defined in Section 8.3.] Use the triangle inequality and induction to prove that for all $n \in \mathbb{N}$ and all $x_1, \ldots, x_n \in V$ we have*

$$\left|\left| \sum_{i=1}^{n} x_i \right|\right| \leq \sum_{i=1}^{n} ||x_i||$$

EXERCISE 8.34. *(Bounded linear maps and continuity) Let U, V be real inner product spaces. Using the inner product norms on U and V, we say that a linear transformation $T : U \to V$ is* bounded *if there is a real-valued constant $\beta \geq 0$ such that*

$$||T(u)|| \leq \beta\, ||u|| \quad \forall u \in U$$

where $||u||$ denotes the inner product norm on U, and $||T(u)||$ uses the inner product norm on V. We say that a function $T : U \to V$ is continuous *if for every $u \in U$ and $\epsilon > 0$, there is a $\delta > 0$ such that*

$$(||u - w|| \leq \delta) \implies ||T(u) - T(w)|| \leq \epsilon$$

a) Prove that all bounded linear transformations $T : U \to V$ are continuous.

b) Prove that if $\dim(U) < \infty$ then every linear transformation $T : U \to V$ is bounded. Hint: The case $U = \{0\}$ is trivial. Assume $\dim(U) = n$ for some $n \in \mathbb{N}$ and consider an orthonormal basis $\{u_1, \ldots, u_n\}$ for U. You are free to use the bounds given in Exercises 8.32 and 8.33.

EXERCISE 8.35. *(A bounded linear transformation) Let U be the vector space of real-valued*

square-summable sequences $(x_1, x_2, \ldots)$. *Define the function* $T : U \to \mathbb{R}$ *by*

$$T(x_1, x_2, x_3, \ldots) = \sum_{i=1}^{\infty} \frac{1}{i} x_i \quad \forall x \in U$$

a) Verify that $T(x) \in \mathbb{R}$ *for all* $x \in U$ *by using* (8.10) *and the fact* $\sum_{i=1}^{\infty} \frac{1}{i^2} < \infty$.

b) Show T *is a linear transformation.*

c) Show T *is a bounded linear transformation, where "bounded" is defined in Exercise 8.34.*

EXERCISE 8.36. *(An unbounded linear transformation) Let* U *be the set of all sequences* $(x_1, x_2, x_3, \ldots) \in \mathbb{R}^{\mathbb{N}}$ *such that* $x_i \in \mathbb{R}$ *for all* $i \in \mathbb{N}$, *and* $x_i = 0$ *for all but finitely many* $i \in \mathbb{N}$. *Define*

$$\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i \quad \forall x, y \in U$$

a) Show that U *is a subspace of* $\mathbb{R}^{\mathbb{N}}$. *[Thus,* U *is a real vector space.]*

b) Show that $\langle x, y \rangle$ *satisfies the requirements of a real inner product.*

c) Define $T : U \to \mathbb{R}$ *by*

$$T(x_1, x_2, x_3, \ldots) = \sum_{i=1}^{\infty} i x_i \quad \forall x \in U$$

where $\sum_{i=1}^{\infty} i x_i \in \mathbb{R}$ *because only a finite number of terms in the sum are nonzero. Show that* T *is a linear transformation.*

d) Show the linear transformation T *is unbounded, where "bounded" is defined in Exercise 8.34.*

EXERCISE 8.37. *(Every basis for* $\mathbb{R}^n$ *is an orthonormal basis) Let* A *be the real* $n \times n$ *matrix from the proof of Theorem 8.2.*

a) Show that A *is symmetric and positive semidefinite.*

b) Show that A *is positive definite. In view of part (a), this amounts to proving that if* x *is a vector in* $\mathbb{R}^n$ *that satisfies* $x^\top A x = 0$, *then* $x = 0$.

EXERCISE 8.38. *(Existence of inner product) Verify that the function* $\langle v, w \rangle = T(v)^\top A T(w)$ *defined in Section 8.10.2 is a valid real inner product on* V.

EXERCISE 8.39. *(All inner products on* V*) Let* V *be a real vector space with* $\dim(V) = n \in \mathbb{N}$. *Let* $\langle v, w \rangle$ *be a real inner product on* V. *We want to show this inner product has the form* (8.29). *Following the development of Section 8.10.2, let* $\{v_1, \ldots, v_n\}$ *be a basis of* V. *Let* $T : V \to \mathbb{R}^n$ *be the invertible linear transformation that maps each* $v \in V$ *to the unique vector* $(x_1, \ldots, x_n) \in \mathbb{R}^n$ *that satisfies* $\sum_{i=1}^{n} x_i v_i = v$. *Define the function* $f : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ *by*

$$f(x, y) = \left\langle T^{-1}(x), T^{-1}(y) \right\rangle \quad \forall x, y \in \mathbb{R}^n$$

a) Show that f is an inner product on $\mathbb{R}^n$. You may want to use properties of invertible linear transformations in Chapter 6.13.

b) Use Lemma 8.4 to conclude that $f(x, y) = x^\top A y$ for all $x, y \in \mathbb{R}^n$, where A is some real symmetric and positive definite matrix of size $n \times n$.

c) Conclude that $\langle v, w \rangle = T(v)^\top A T(w)$ for all $v, w \in V$.

EXERCISE 8.40. *Let V be a real inner product space. Let R and S be subspaces such that $R \subseteq S$. Show that $R^\perp \supseteq S^\perp$.*

EXERCISE 8.41. *(Parallelogram law implies inner product norm) Let V be a real vector space. Suppose $||\cdot||$ is a norm on V that satisfies the parallelogram law. Define $f : V \times V \to \mathbb{R}$ by*

$$f(x, y) = \frac{||x + y||^2 - ||x - y||^2}{4} \quad \forall x, y \in V$$

We want to show f is an inner product on V, and $||\cdot||$ is the inner product norm with respect to this inner product. [The structure of function f is motivated by the polarization identity, which is the only possible form that the inner product can take if $||\cdot||$ is in fact an inner product norm.]

a) Nonnegativity: Show that $f(x, x) = ||x||^2$ for all $x \in V$. Conclude that f satisfies the nonnegativity property of inner products.

b) Symmetry: Show that $f(x, y) = f(y, x)$ for all $x, y \in V$.

c) Fix $x, y, z \in V$. Use the fact that $||\cdot||$ satisfies the parallelogram law to prove

$$||x + y + z||^2 = 2\,||x + z||^2 + 2\,||y||^2 - ||x + z - y||^2$$
$$||x + y - z||^2 = 2\,||x - z||^2 + 2\,||y||^2 - ||x - z - y||^2$$

Conclude that

$$f(x + y, z) = 2f(x, z) + \frac{||x - z - y||^2 - ||x + z - y||^2}{4}$$

d) By switching roles of x and y, argue that

$$f(x + y, z) = 2f(y, z) + \frac{||y - z - x||^2 - ||y + z - x||^2}{4}$$

Add the two equalities for $f(x+y, z)$ in (c) and (d) to conclude $f(x+y, z) = f(x, z) + f(y, z)$. By symmetry, conclude that $f(z, x+y) = f(z, x) + f(z, y)$, so f satisfies the summation property in the second slot. [Note: While not part of this exercise, we briefly show how to conclude f is linear in the second slot. Exercise 6.55 shows this summation property implies $f(x, ay) = af(x, y)$ for all $x, y \in V$ and all rational numbers *$a \in \mathbb{Q}$. Also, for each fixed $x \in V$, $f(x, y)$ is a continuous function of $y \in V$ (see Exercise 8.42), so $f(x, ay) = af(x, y)$ for all $a \in \mathbb{R}$ and $y \in V$.]*

EXERCISE 8.42. *(Continuity of norm) Let V be a real vector space. Let $||\cdot||$ be a norm on V.*

a) Use the triangle inequality to show that for all $x, y \in V$ we have

$$||x|| - ||y|| \leq ||x + y|| \leq ||x|| + ||y||$$

[If a hint is desired, it is in a famous bullet of Lemma 6.1.]

b) Suppose $\{x, h_1, h_2, h_3, ...\}$ are vectors in V and $\lim_{n\to\infty} ||h_n|| = 0$. Show that

$$\lim_{n\to\infty} ||x + h_n|| = ||x||$$

[This implies for any continuous function $g : \mathbb{R} \to \mathbb{R}$ we have $\lim_{n\to\infty} g(||x + h_n||) = g(||x||)$.]

EXERCISE 8.43. *(Manufacturing functions in $S^\perp$) Let V be the real vector space of square integrable functions $f : [0, 1] \to \mathbb{R}$. Use the standard inner product $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$ for $f, g \in V$. Define $S = Span\{t, e^t\}$. Lemma 8.9(a) ensures*

$$S^\perp = \left\{ f \in V : \int_0^1 f(t)tdt = 0 \text{ and } \int_0^1 f(t)e^t dt = 0 \right\}$$

We construct some example functions $f \in S^\perp$. Let g be any function in V. Define $f : [0, 1] \to \mathbb{R}$ by

$$f(t) = \begin{cases} g(t) & \text{if } t \in [0, 1/2) \\ a + bt & \text{if } t \in [1/2, 1] \end{cases}$$

where a, b are some constants in $\mathbb{R}$.

a) Show that for all $a, b \in \mathbb{R}$ we have $f \in V$.

b) Design a, b such that $f \in S^\perp$.

CHAPTER 9.

Projections

Let V be a real inner product space with inner product $\langle x, y \rangle$. Let $||x|| = \sqrt{\langle x, x \rangle}$ be the inner product norm. The norm can be used to define a *distance* between two vectors in V:

$$\text{distance}(x, y) = ||x - y|| \quad \forall x, y \in V$$

Then $||x||$ is the distance between x and the zero vector:

$$||x|| = ||x - 0||$$

The value $||x||^2$ is often associated with the *energy* in vector x. Depending on the structure of V, a vector $x \in V$ might be a n-tuple, a function, a random variable, and so on.

Let S be a subspace of V. For a given $y \in V$, we want to find a vector in S that is closest to y (if such a closest point exists). That is, we want to find p that solves

$$\begin{aligned} &\text{Minimize:} && ||y - p|| \\ &\text{Subject to:} && p \in S \end{aligned}$$

Minimizing $||y - p||$ is the same as minimizing $||y - p||^2$, so an equivalent problem is

$$\text{Minimize:} \quad ||y - p||^2 \tag{9.1}$$

$$\text{Subject to:} \quad p \in S \tag{9.2}$$

If $y \in S$ then clearly the unique answer is $p = y$. If $y \notin S$, we want a vector $p \in S$ that minimizes the squared error $||y - p||^2$.

For example, suppose V is the real vector space of functions $f : [0,1] \to \mathbb{R}$ that are square integrable. Use the standard inner product on this space. Let $S = Span\{e^t, t-1\}$. Given $y \in V$, the problem (9.1)-(9.2) becomes

$$\begin{aligned} \text{Minimize:} \quad & \int_0^1 (y(t) - p(t))^2 dt \\ \text{Subject to:} \quad & p \in Span\{e^t, t-1\} \end{aligned}$$

Does a solution p exist? Is it unique? If so, how can a solution be found? Problems of this type are related to function approximation and Fourier series.

For another example, suppose we have a real $m \times n$ matrix A and a vector $y \in \mathbb{R}^m$. We want to find $x \in \mathbb{R}^n$ such that $y \approx Ax$. To make the approximation goal precise, consider the problem of finding x to solve

$$\text{Minimize:} \quad ||y - Ax||^2 \tag{9.3}$$

$$\text{Subject to:} \quad x \in \mathbb{R}^n \tag{9.4}$$

where we use the standard Euclidean norm. If there exists a vector x that satisfies $y = Ax$ with exact equality, then x solves (9.3)-(9.4). However, it is often impossible to solve $y = Ax$ exactly, so we settle for a vector x that minimizes the squared error $||y - Ax||^2$. The problem (9.3)-(9.4) is equivalent to

$$\begin{aligned} \text{Minimize:} \quad & ||y - p||^2 \\ \text{Subject to:} \quad & p \in Col(A) \end{aligned}$$

which has the form (9.1)-(9.2) with $V = \mathbb{R}^m$ and $S = Col(A)$. Problems of this type are called *least squares* or *linear regression* problems. They are often used to build models that best fit a collection of data.

This chapter shows that if the general problem (9.1)-(9.2) has a solution, then: (i) the solution is unique; (ii) it can be found by taking the *orthogonal projection of y onto subspace S*. Further, problem (9.1)-(9.2) is guaranteed to have a solution when $\dim(S) < \infty$.

9.1. Closest point counter-examples

Consider the problem of finding a point in a given subset $S \subseteq \mathbb{R}^2$ that is closest to a given $y = (y_1, y_2) \in \mathbb{R}^2$. A closest point is a vector $p = (p_1, p_2) \in S$ that satisfies

$$||y - p||^2 \leq ||y - s||^2 \quad \forall s \in S$$

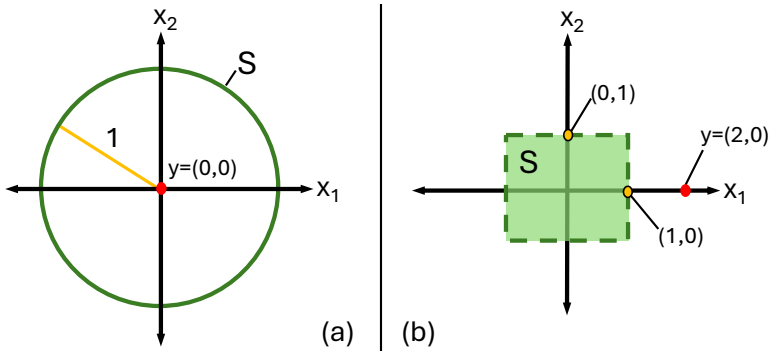


FIGURE 9.1. (a) The circle S for Example 1; (b) The open box S for Example 2.

where we use the standard inner product norm for $\mathbb{R}^2$. Equivalently

$$(y_1 - p_1)^2 + (y_2 - p_2)^2 \leq (y_1 - s_1)^2 + (y_2 - s_2)^2 \quad \forall (s_1, s_2) \in S$$

In the following examples, the set S is *not* a subspace, which is why existence and uniqueness can fail.

- **Example 1:** $y = (0, 0)$ and $S = \{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ (see Fig. 9.1a). The set S is a circle of radius 1 about the origin. Is there a closest point in S to y? Yes: Every point in S is a distance 1 from y. There are infinitely many closest points. *The closest points are not unique.*
- **Example 2:** $y = (2, 0)$ and $S = \{(x_1, x_2) \in \mathbb{R}^2 : -1 < x_1 < 1, -1 < x_2 < 1\}$ (see Fig. 9.1b). The set S is an open box. Is there a closest point in S to y? No. [Note that $(1, 0)$ is on the boundary of S, but is not in S itself.] For any $(x_1, x_2) \in S$, the reader can verify that $(\frac{x_1+1}{2}, 0)$ is also in S, and is closer to y. So (x_1, x_2) fails to be a closest point. *A closest point does not exist.*

9.2. Uniqueness of closest point in a subspace

THEOREM 9.1. *Let V be a real inner product space. Define distances by the inner product norm. Let S be a subspace of V. Fix $y \in V$. Suppose there exists a point p that is closest in S to the point y. That is:*

1. $p \in S$
2. $||y - p|| \leq ||y - s||$ *for all* $s \in S$.

Then p is the unique *closest point in S to y and is called the* orthogonal projection of y onto the set S. *In particular, if $s \in S$ and $s \neq p$, then $||y - p|| < ||y - s||$.*

PROOF. Suppose p and s are two distinct closest points in S to y (we reach a contradic-

tion). We know $s \neq p$ and

$$\|y - p\| = \|y - s\| \tag{9.5}$$

We reach a contradiction by constructing a point in S that is closer to y than p is. We use the strong convexity equality of inner product norms (Lemma 8.3). For all $t \in \mathbb{R}$ we have

$$\|t(y-s) + (1-t)(y-p)\|^2 = t\|y-s\|^2 + (1-t)\|y-p\|^2 - t(1-t)\|(y-s)-(y-p)\|^2$$

Simplifying the above equality and using $t = 1/2$ gives

$$\begin{aligned}
\left\|y - \frac{s+p}{2}\right\|^2 &= \frac{1}{2}\|y-s\|^2 + \frac{1}{2}\|y-p\|^2 - \frac{1}{4}\|p-s\|^2 \\
&\overset{(a)}{<} \frac{1}{2}\|y-s\|^2 + \frac{1}{2}\|y-p\|^2 \\
&\overset{(b)}{=} \frac{1}{2}\|y-p\|^2 + \frac{1}{2}\|y-p\|^2 \\
&= \|y-p\|^2
\end{aligned}$$

where the strict inequality (a) follows because $s \neq p$, so $\|p-s\|^2 > 0$; equality (b) holds by (9.5). Notice that $\frac{s+p}{2} \in S$ because it is a linear combination of the vectors $s, p \in S$. We conclude that $\frac{s+p}{2}$ is a vector in S that is *strictly closer to y* than p is, which contradicts our assumption that p is a closest point in S to y. □

9.3. Existence of closest point in a finite dimensional subspace

THEOREM 9.2. *Let V be a real inner product space. Define distances by the inner product norm. Let S be a subspace of V. Suppose* $\dim(S) < \infty$. *Then for every $y \in V$, there is a point $p \in S$ that is the closest in S to y.*

PROOF. If $\dim(S) = 0$ then $S = \{0\}$ and the result is trivial. Suppose $\dim(S) = n$ for some $n \in \mathbb{N}$. Then S has an orthonormal basis $\{v_1, \ldots, v_n\}$. Fix $y \in V$. Define $f : \mathbb{R}^n \to \mathbb{R}$ by

$$f(\theta) = \left\|y - \textstyle\sum_{i=1}^n \theta_i v_i\right\|^2 \quad \forall \theta \in \mathbb{R}^n$$

The problem of finding a point $p \in S$ that is closest to y is equivalent to finding $\theta \in \mathbb{R}^n$ to minimize f. The function f can be written:

$$\begin{aligned}
f(\theta) &\overset{(a)}{=} \|y\|^2 + \left\|\textstyle\sum_{i=1}^n \theta_i v_i\right\|^2 - 2\left\langle y, \textstyle\sum_{i=1}^n \theta_i v_i\right\rangle \\
&\overset{(b)}{=} \|y\|^2 + \sum_{i=1}^n \theta_i^2 - 2\sum_{i=1}^n \theta_i \langle y, v_i\rangle
\end{aligned}$$

$$= \|y\|^2 + \sum_{i=1}^{n} \left[\theta_i^2 - 2\theta_i \langle y, v_i \rangle \right]$$

$$= \|y\|^2 - \sum_{i=1}^{n} \langle y, v_i \rangle^2 + \sum_{i=1}^{n} (\theta_i - \langle y, v_i \rangle)^2 \tag{9.6}$$

where (a) uses $\|a+b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b \rangle$; (b) uses orthonormality of $\{v_1, ..., v_n\}$ and Lemma 8.6b. The only term on the right-hand-side of (9.6) that depends on θ is the last one. Clearly f has minimizer $\theta^* = (\theta_1^*, ..., \theta_n^*)$ where

$$\theta_i^* = \langle y, v_i \rangle \quad \forall i \in \{1, ..., n\}$$

In particular, the closest point in S to y is the point $p = \sum_{i=1}^{n} \langle y, v_i \rangle v_i$. □

This proof of the existence theorem shows that an explicit closest point can be computed when S has finite dimension and a known orthonormal basis. When the existence and uniqueness theorems are taken together, they show that if V is any (possibly infinite-dimensional) real inner product space, and if S is any finite-dimensional subspace of V, then every $y \in V$ has a unique closest point in the subspace S. We can use the notation $p = Proj_S(y)$ to mean that p is the orthogonal projection of vector y onto the subspace S.

9.3.1. Terminology

Throughout this chapter, we use the term *projection* to mean the orthogonal projection, where distances and orthogonality are with respect to the particular inner product of interest. Section 9.16 presents a more general definition of *projection operator*, then shows every projection operator on a real vector space is an orthogonal projection onto some subspace and under some inner product.

9.3.2. An infinite dimensional counter-example

If subspace S is infinite dimensional then existence of a closest point can fail. For example, suppose V is the inner product space of square integrable functions $f : [0, 1] \to \mathbb{R}$. Let S be the subspace of polynomial functions (of any degree). This is a subspace because the sum of two polynomials is again a polynomial, as is a scalar multiple of a polynomial. We have $\dim(S) = \infty$ because for any positive integer k, the functions $\{1, t, t^2, ..., t^k\}$ are linearly independent in S. The function $e^t = 1 + \sum_{n=1}^{\infty} \frac{t^n}{n!}$ is in V but there is no closest point in S. Indeed, for any particular polynomial function p, there is another polynomial function q that is closer to e^t, that is

$$0 < \left\| q - e^t \right\| < \left\| p - e^t \right\|$$

See the related Exercise 6.29.

9.4. Orthogonal projection theorem

The following theorem holds for any (possibly infinite-dimensional) real inner product space V and any (possibly infinite-dimensional) subspace $S \subseteq V$. It does not require the existence theorem (Theorem 9.2) because it concerns a particular situation.

THEOREM 9.3. *(Orthogonal projection theorem) Suppose V is a real inner product space. Define distances by the inner product norm. Let S be a subspace of V. Fix $y \in V$. A vector p is the closest point in S to y if and only if $p \in S$ and the following* orthogonality conditions *hold:*

$$\langle y - p, s \rangle = 0 \quad \forall s \in S \tag{9.7}$$

(The condition (9.7) *means $y - p \in S^\perp$.) In particular, if $S = Span(C)$ for some collection of vectors $C \subseteq V$, then p is the closest point in S to y if and only if $p \in S$ and $\langle y - p, v \rangle = 0$ for all $v \in C$.*

PROOF. First suppose that $p \in S$ and the orthogonality conditions hold. We want to show that p is the closest point in S to y. Fix $s \in S$. It suffices to show $||y - s||^2 \geq ||y - p||^2$. We have

$$\begin{aligned} ||y - s||^2 &= ||(y - p) + (p - s)||^2 \\ &= ||y - p||^2 + ||p - s||^2 + 2\langle y - p, p - s \rangle \\ &\overset{(a)}{=} ||y - p||^2 + ||p - s||^2 \\ &\geq ||y - p||^2 \end{aligned}$$

where (a) uses the fact that $\langle y - p, p - s \rangle = 0$ (since $p - s \in S$ and the orthogonality conditions hold). This completes the proof of the reverse direction.

To prove the forward direction, suppose $p \in S$ and p is the closest point in S to y. We want to show the orthogonality conditions hold. Fix $s \in S$. It suffices to show $\langle y - p, s \rangle = 0$. Fix $\epsilon \in \mathbb{R}$. Since $p \in S$ and $s \in S$ we know $p + \epsilon s \in S$ (recall that S is a subspace). Since p is the closest point in S to y we know

$$||y - p||^2 \leq ||y - (p + \epsilon s)||^2$$

Expanding the right-hand-side gives

$$||y - (p + \epsilon s)||^2 = ||(y - p) - \epsilon s||^2 = ||y - p||^2 + \epsilon^2 ||s||^2 - 2\epsilon \langle y - p, s \rangle$$

Substituting this into the previous inequality and simplifying gives

$$0 \leq \epsilon^2 ||s||^2 - 2\epsilon \langle y - p, s \rangle \tag{9.8}$$

The inequality (9.8) holds for all $\epsilon \in \mathbb{R}$. Suppose $\epsilon > 0$. Dividing both sides of (9.8) by 2ϵ gives

$$0 \leq \frac{\epsilon}{2} ||s||^2 - \langle y - p, s \rangle \quad \forall \epsilon > 0$$

Taking $\epsilon \searrow 0$ gives

$$0 \leq - \langle y - p, s \rangle \tag{9.9}$$

Now suppose $\epsilon < 0$. Dividing both sides of (9.8) by 2ϵ gives

$$0 \geq \frac{\epsilon}{2} ||s||^2 - \langle y - p, s \rangle \quad \forall \epsilon < 0$$

where the inequality has been flipped since we have divided both sides by a negative number. Taking $\epsilon \nearrow 0$ gives

$$0 \geq - \langle y - p, s \rangle$$

Combining this with (9.9) gives $\langle y - p, s \rangle = 0$. This proves the main result. The special case when $S = Span(C)$ holds by Lemma 8.9(b). □

9.5. Projecting onto a 1-dimensional subspace

Let V be a real inner product space. Fix v as a nonzero vector in V and define $S = Span\{v\}$. Let y be another vector in V. We want to project y onto $Span\{v\}$. The projection exists and is unique because $Span\{v\}$ is a finite-dimensional subspace of V (it has only 1 dimension). Let p be the projection of y onto $Span\{v\}$. The orthogonality condition is

$$\langle y - p, s \rangle = 0 \quad \forall s \in Span\{v\}$$

and since $v \in S$, this implies $\langle y - p, v \rangle = 0$. Thus

$$\langle y, v \rangle = \langle p, v \rangle \tag{9.10}$$

Since $p \in Span\{v\}$ we know $p = av$ for some $a \in \mathbb{R}$. Substituting $p = av$ into (9.10) gives

$$\langle y, v \rangle = \langle av, v \rangle = a ||v||^2$$

Since $v \neq 0$, we conclude

$$a = \frac{\langle y, v \rangle}{||v||^2} \tag{9.11}$$

$$p = \frac{\langle y, v \rangle v}{||v||^2} \tag{9.12}$$

It is sometimes useful to define q as a normalized version of v:

$$q = \frac{v}{||v||} \qquad [\text{so } ||q|| = 1]$$

Then $S = Span\{v\} = Span\{q\}$. Using (9.12), we can equivalently express p in terms of q:

$$p = \frac{\langle y, v\rangle v}{||v||^2} = \left\langle y, \frac{v}{||v||}\right\rangle \frac{v}{||v||} = \langle y, q\rangle q$$

The value $\langle y, q\rangle$ is a real number that can be viewed as the *component of y that is in the direction of q* (the direction of q is the same as v). The orthogonality condition ensures that $y - p$ is orthogonal to everything in the direction of q. In particular, $y - p$ is orthogonal to p (see also Exercise 9.15).

We obtained the projection p from the orthogonality condition. Of course, p can also be obtained by noting $p = \theta v$ for some $\theta \in \mathbb{R}$, then finding a value $\theta \in \mathbb{R}$ to solve

$$\begin{aligned} &\text{Minimize:} && ||y - \theta v||^2 \\ &\text{Subject to:} && \theta \in \mathbb{R} \end{aligned}$$

The minimization objective is

$$||y - \theta v||^2 = \langle y - \theta v, y - \theta v\rangle = ||y||^2 + \theta^2 ||v||^2 - 2\theta \langle y, v\rangle$$

and since $||v||^2 > 0$, this is a convex parabola in θ. The optimal θ^* is found by setting the derivative to 0:

$$0 = \frac{d}{d\theta}\left[||y||^2 + \theta^2 ||v||^2 - 2\theta \langle y, v\rangle\right] = 2\theta ||v||^2 - 2\langle y, v\rangle$$

which yields $\theta^* = \frac{\langle y,v\rangle}{||v||^2}$ and $p = \theta^* v$, so the optimal θ^* is the same as the value a in (9.11).

9.6. Approximating e^t

Define V as the vector space of square integrable functions $f : [0, 1] \to \mathbb{R}$. We use inner product

$$\langle f, g\rangle = \int_0^1 f(t)g(t)dt \quad \forall f, g \in V$$

and the concept of equivalence classes, so $f \equiv g$ if and only if $\int_0^1 (f(t) - g(t))^2 dt = 0$. Let S be the subspace of polynomial functions of degree less than or equal to 2. Subspace S has basis $\{1, t, t^2\}$. We want to project the function e^t onto S. Equivalently, we want the best approximation of e^t of the form $a_0 + a_1 t + a_2 t^2$, where "best" means that we choose real

numbers a_0, a_1, a_2 to minimize

$$\left\| e^t - \left(\sum_{i=0}^{2} a_i t^i \right) \right\|^2 = \int_0^1 \left(e^t - \sum_{i=0}^{2} a_i t^i \right)^2 dt$$

One quadratic approximation of e^t is given by the Taylor expansion about $t = 0$:

$$e^t \approx 1 + t + \frac{t^2}{2}$$

Of course, this Taylor expansion does not necessarily minimize the squared integral of the error across $[0, 1]$. It is interesting to compare the optimal projection to the Taylor expansion. We first find the projection.

Since S is a 3-dimensional subspace of V, we know a projection exists and is unique. Let $p(t) \in S$ be the projection. By the projection theorem, we know $p(t)$ satisfies the orthogonality condition:

$$\left\langle e^t - p, s \right\rangle = 0 \quad \forall s \in S$$

Since $\{1, t, t^2\}$ is a basis for S, we know that

$$\begin{aligned} \left\langle e^t - p, 1 \right\rangle &= 0 \\ \left\langle e^t - p, t \right\rangle &= 0 \\ \left\langle e^t - p, t^2 \right\rangle &= 0 \end{aligned}$$

Since $p(t) \in S$, it has the form $p(t) = a_0 + a_1 t + a_2 t^2$ for some real numbers a_i, and so

$$\begin{aligned} \left\langle e^t - a_0 - a_1 t - a_2 t^2, 1 \right\rangle = 0 &\implies \left\langle e^t, 1 \right\rangle = a_0 \langle 1, 1 \rangle + a_1 \langle t, 1 \rangle + a_2 \left\langle t^2, 1 \right\rangle \\ \left\langle e^t - a_0 - a_1 t - a_2 t^2, t \right\rangle = 0 &\implies \left\langle e^t, t \right\rangle = a_0 \langle 1, t \rangle + a_1 \langle t, t \rangle + a_2 \left\langle t^2, t \right\rangle \\ \left\langle e^t - a_0 - a_1 t - a_2 t^2, t^2 \right\rangle = 0 &\implies \left\langle e^t, t^2 \right\rangle = a_0 \left\langle 1, t^2 \right\rangle + a_1 \left\langle t, t^2 \right\rangle + a_2 \left\langle t^2, t^2 \right\rangle \end{aligned}$$

By symmetry of the inner product we know $\langle t, 1 \rangle = \langle 1, t \rangle$ and so on, so the above system of equations can be equally written

$$\begin{aligned} \left\langle e^t, 1 \right\rangle &= a_0 \langle 1, 1 \rangle + a_1 \langle 1, t \rangle + a_2 \left\langle 1, t^2 \right\rangle \\ \left\langle e^t, t \right\rangle &= a_0 \langle t, 1 \rangle + a_1 \langle t, t \rangle + a_2 \left\langle t, t^2 \right\rangle \\ \left\langle e^t, t^2 \right\rangle &= a_0 \left\langle t^2, 1 \right\rangle + a_1 \left\langle t^2, t \right\rangle + a_2 \left\langle t^2, t^2 \right\rangle \end{aligned}$$

which has the matrix form

$$\underbrace{\begin{bmatrix} \langle e^t, 1\rangle \\ \langle e^t, t\rangle \\ \langle e^t, t^2\rangle \end{bmatrix}}_{h} = \underbrace{\begin{bmatrix} \langle 1,1\rangle & \langle 1,t\rangle & \langle 1,t^2\rangle \\ \langle t,1\rangle & \langle t,t\rangle & \langle t,t^2\rangle \\ \langle t^2,1\rangle & \langle t^2,t\rangle & \langle t^2,t^2\rangle \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}}_{a}$$

so $h = Ga$, where $G = (G_{ij})$ is the Gram matrix for $\{1, t, t^2\}$, and $h = (h_1, h_2, h_3)$ is the vector of inner products between e^t and $\{1, t, t^2\}$:

$$\begin{aligned} h_1 &= \langle e^t, 1\rangle = \int_0^1 e^t dt = e - 1 \\ h_2 &= \langle e^t, t\rangle = \int_0^1 te^t dt = 1 \\ h_3 &= \langle e^t, t^2\rangle = \int_0^1 t^2 e^t dt = e - 2 \\ G_{11} &= \langle 1, 1\rangle = \int_0^1 1 dt = 1 \\ G_{12} &= \langle 1, t\rangle = \int_0^1 t dt = 1/2 \\ G_{13} &= \langle 1, t^2\rangle = \int_0^1 t^2 dt = 1/3 \\ G_{22} &= \langle t, t\rangle = \int_0^1 t^2 dt = 1/3 \\ G_{23} &= \langle t, t^2\rangle = \int_0^1 t^3 dt = 1/4 \\ G_{33} &= \langle t^2, t^2\rangle = \int_0^1 t^4 dt = 1/5 \end{aligned}$$

The $h = Ga$ equation becomes

$$\begin{bmatrix} e-1 \\ 1 \\ e-2 \end{bmatrix} = \begin{bmatrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \end{bmatrix}$$

Gram matrix G is invertible because $\{1, t, t^2\}$ is linearly independent (Lemma 8.13). Then $a = G^{-1}h$ and

$$\begin{aligned} a_0 &= 1.012991309902759 \\ a_1 &= 0.851125052846271 \\ a_2 &= 0.839183976399454 \end{aligned}$$

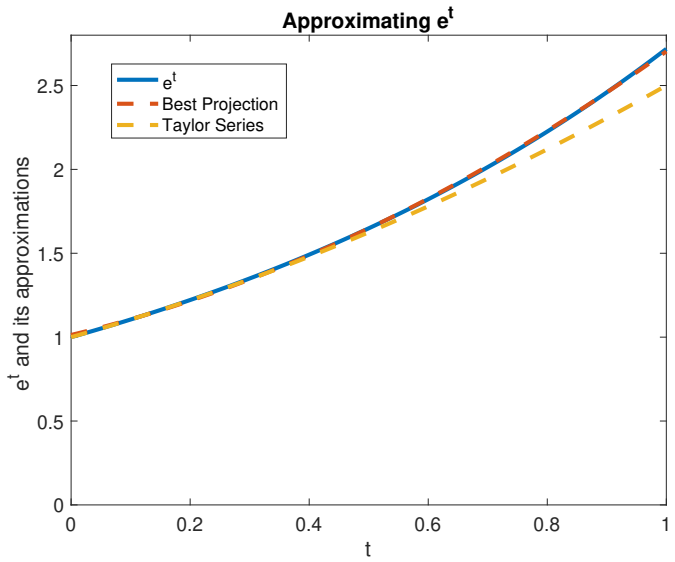


FIGURE 9.2. Function e^t over domain $[0,1]$ compared with its projection $p(t) = a_0 + a_1 t + a_2 t^2$ onto the space of polynomials of degree ≤ 2, and the Taylor expansion $\tilde{p}(t) = 1+t+t^2/2$. The projection $p(t)$ is more accurate over the full interval $[0,1]$.

Thus, the projection is $p(t) = a_0 + a_1 t + a_2 t^2$ with numerical values of a_i given above. For comparison, the Taylor expansion $\tilde{p}(t) = 1 + t + t^2/2$ uses weights $\tilde{a}_0 = 1$, $\tilde{a}_1 = 1$, $\tilde{a}_2 = 1/2$. So the optimal projection $p(t)$ weights the constant term slightly more in comparison to Taylor, the linear term less, and the quadratic term more. Note that Taylor is exact at $t = 0$ because $\tilde{p}(t) = 1 = e^0$, while the projection is inexact at $t = 0$ because

$$p(0) = a_0 = 1.012991309902759$$

Fig. 9.2 compares e^t with its projection $p(t) = a_0 + a_1 t + a_2 t^2$ (using the numerical coefficients given above) and with its Taylor expansion $1 + t + t^2/2$. While the Taylor expansion is accurate for t near 0, it is significantly less accurate over the full interval $[0,1]$ in comparison to the projection $p(t)$.

If we wanted to improve accuracy for small values of t in the interval $[0,1]$ we could use an inner product that weights those values more heavily, such as

$$\langle f, g \rangle = \int_0^1 e^{1-t} f(t) g(t) dt \quad \forall f, g \in V$$

which can be shown to be another inner product on V. To project e^t onto S using this inner product, we use $p(t) = a_0 + a_1 t + a_2 t^2$ with $a = (a_0, a_1, a_2)$ again found by solving $h = Ga$, but the Gram matrix G is different (as is the h vector). See also Exercise 9.13.

9.7. Projecting onto $Col(A)$

Let A be a $m \times n$ matrix with entries in $\mathbb{R}$. Fix $y \in \mathbb{R}^m$. Suppose we want to find a vector $x \in \mathbb{R}^n$ that satisfies

$$y \approx Ax$$

Specifically, we want to find an x that solves

$$\begin{aligned} \text{Minimize:} \quad & \|y - Ax\|^2 \\ \text{Subject to:} \quad & x \in \mathbb{R}^n \end{aligned}$$

which uses the standard inner product $\langle u, v \rangle = u^\top v$ and the corresponding norm $\|u\| = \sqrt{\langle u, u \rangle}$. This is closely related to the problem of projecting y onto $Col(A)$:

$$\begin{aligned} \text{Minimize:} \quad & \|y - p\|^2 \\ \text{Subject to:} \quad & p \in Col(A) \end{aligned}$$

Since $Col(A)$ is finite-dimensional, we know a unique solution to this problem exists. Let $p^* \in Col(A)$ be the unique solution, being the unique vector in $Col(A)$ that is closest to y. Then any point $x \in \mathbb{R}^n$ that satisfies $Ax = p^*$ is a solution to the first problem. Since $p^* \in Col(A)$, we know there is *at least one particular* $x^* \in \mathbb{R}^n$ that satisfies $Ax^* = p^*$. Whether or not x^* is unique depends on how many vectors $Null(A)$ contains.

LEMMA 9.1. *Let A be a real $m \times n$ matrix and fix $y \in \mathbb{R}^m$. Then*

a) The following equation has at least one solution $x \in \mathbb{R}^n$:

$$A^\top Ax = A^\top y \tag{9.13}$$

Further, if x_1 and x_2 are any two solutions to this equation, then $Ax_1 = Ax_2$.

b) If x^* is any particular *solution to $A^\top Ax = A^\top y$, then the following vector p^* is the unique projection of y onto $Col(A)$:*

$$p^* = Ax^*$$

c) If $Null(A) = \{0\}$ (which requires A to be square or tall) then $A^\top A$ is invertible and the equation of part (a) has one unique solution. In particular:[1]

$$x^* = (A^\top A)^{-1} A^\top y \tag{9.14}$$

$$p^* = A(A^\top A)^{-1} A^\top y \tag{9.15}$$

[1] For a general real matrix A of size $m \times n$, it holds that $Null(A) = Null(A^\top A)$ and so the square matrix $A^\top A$ is invertible if and only if $Null(A) = \{0\}$.

PROOF. To prove (a), let us project y onto $Col(A)$. Since $Col(A)$ has finite dimension, we know there exists a unique projection $p \in Col(A)$. The projection must satisfy the orthogonality conditions:

$$(y - p)^\top s = 0 \quad \forall s \in Col(A)$$

Since $p \in Col(A)$ we know $p = Ax$ for some $x \in \mathbb{R}^n$. The above conditions become:

$$(y - Ax)^\top s = 0 \quad \forall s \in Col(A)$$

Since $s \in Col(A)$ if and only if $s = Aw$ for some w, we know

$$(y - Ax)^\top Aw = 0 \quad \forall w \in \mathbb{R}^n$$

Taking transposes gives

$$w^\top A^\top (y - Ax) = 0 \quad \forall w \in \mathbb{R}^n$$

Then vector $A^\top (y - Ax)$ is orthogonal to all vectors in $\mathbb{R}^n$, which means it is orthogonal to itself, so it must be zero:

$$A^\top (y - Ax) = 0$$

This means $A^\top y = A^\top Ax$, proving that (9.13) has a solution. Now let x_1 and x_2 be two solutions to $A^\top Ax = A^\top y$. Then $x_1 - x_2 \in Null(A^\top A) = Null(A)$ (recall Lemma 8.5). So $A(x_1 - x_2) = 0$, meaning that $Ax_1 = Ax_2$.

The proof of (a) has essentially also proven (b): We know a projection $p^* \in Col(A)$ exists and satisfies $p^* = Ax^*$ for some $x^* \in \mathbb{R}^n$ that satisfies the orthogonality equations, so part (a) implies $A^\top Ax^* = A^\top y$. But part (a) also shows that *any* solution x to $A^\top Ax = A^\top y$ yields $Ax = Ax^* = p^*$.

To prove (c), let x^* be any particular solution to $A^\top Ax = A^\top y$. So

$$A^\top Ax^* = A^\top y$$

We know $Null(A) = Null(A^\top A)$ (recall Lemma 8.5). If $Null(A) = \{0\}$ then the square matrix $A^\top A$ has a trivial null space, so it is invertible. Multiplying the above by $(A^\top A)^{-1}$ gives the expression for x^* and shows x^* is unique. □

9.7.1. Projection matrix

Notice that $A^\top A$ is the Gram matrix for the vectors $\{v_1, ..., v_n\}$ formed by the columns of A. This is another reason that $A^\top A$ is invertible if and only if the columns of A are linearly

independent (recall Lemma 8.13). If A has linearly independent columns then the matrix

$$A(A^\top A)^{-1}A^\top$$

is called the *projection matrix* (for projecting y onto $Col(A)$). This is because the following linear function $P(y)$, defined for all $y \in \mathbb{R}^m$, projects y onto $Col(A)$:

$$P(y) = A(A^\top A)^{-1}A^\top y \quad \forall y \in \mathbb{R}^m \tag{9.16}$$

In particular, if p_1 is the projection of y_1 onto $Col(A)$, and p_2 is the projection of y_2 onto $Col(A)$, then $p_1 + p_2$ is the projection of $y_1 + y_2$ onto $Col(A)$. More generally, using the orthogonality conditions, the projection operation can be proven to be a linear function of y even if A has linearly dependent columns (see Exercise 9.5). The reader is invited to show $[A(A^\top A)^{-1}A^\top]^2 = [A(A^\top A)^{-1}A^\top]$, and the function (9.16) satisfies

$$(y \in Col(A)) \iff (P(y) = y)$$

$$P(P(y)) = P(y) \quad \forall y \in \mathbb{R}^m$$

9.7.2. Example

For a simple example, suppose

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 0 \\ 1 & 2 \\ 1 & 8 \end{bmatrix} \quad , \quad y = \begin{bmatrix} 2 \\ 2 \\ -1 \\ 1 \end{bmatrix} \tag{9.17}$$

The matrix A has linearly independent columns so $A^\top A$ is invertible. To find $x \in \mathbb{R}^n$ that satisfies, $y \approx Ax$ we have

$$x^* = (A^\top A)^{-1}A^\top y = \begin{bmatrix} 0.581395348837210 \\ 0.023255813953488 \end{bmatrix} \quad , \quad Ax^* = \begin{bmatrix} 0.604651162790698 \\ 0 \\ 0.627906976744186 \\ 0.767441860465116 \end{bmatrix} \approx \begin{bmatrix} 2 \\ 2 \\ -1 \\ 1 \end{bmatrix}$$

One might object by noticing that:

- 0 is *not* approximately 2.
- 0.627906976744186 is *not* approximately −1.

However, this is the best we can do if we want to minimize the sum of squared error in each component. If we want to improve accuracy in any particular component then we can use a different inner product (similar to the *weighted* least squares of Section 9.9). Of course, since both columns of A have a 0 in row 2, any linear combination of those columns *must* produce a 0 in row 2. Therefore, even weighted least squares cannot improve accuracy in the second row. It is possible to choose $x \in \mathbb{R}^2$ so that Ax has *perfect accuracy* in rows 1 and 3 (while sacrificing accuracy in row 4), but the result has a considerably higher sum squared error than the solution x^* given above (see Exercise 9.16). Indeed, since x^* is optimal, we know

$$\left\|y - Ax^*\right\|^2 \leq \|y - Ax\|^2 \quad \forall x \in \mathbb{R}^2$$

This example uses a real matrix A with linearly independent columns, so $A^\top A$ is invertible. If A had linearly *dependent* columns, we could either throw away the redundant columns to produce a thinner matrix (without changing the column space), or we could find any particular solution to $A^\top Ax = A^\top y$.

9.8. Least squares

Suppose we have a system that takes a vector input $(u, v) \in \mathbb{R}^2$ and produces a real-valued output y. Specifically, we can use the system repeatedly, and on the ith usage, the input $(u_i, v_i) \in \mathbb{R}^2$ yields the following output:

$$y_i = c_0 + c_1 u_i + c_2 v_i + c_3 u_i v_i + N_i \tag{9.18}$$

where c_0, c_1, c_2, c_3 are unknown real-valued coefficients, and N_i is an unknown random disturbance. The random disturbance is assumed to be "centered about 0" in some way, such as independent random noise with zero mean (so $\mathbb{E}[N_i] = 0$ for all i).

To estimate the unknown coefficients c_0, c_1, c_2, c_3 we take m measurements to obtain triplets (u_i, v_i, y_i) for $i \in \{1, \ldots, m\}$. The triplets can be viewed as known data. Using this data, we build coefficients (a_0, a_1, a_2, a_3) that are *estimates* of the true coefficients (c_0, c_1, c_2, c_3). How can we do this? One reasonable goal is to find $(a_0, a_1, a_2, a_3) \in \mathbb{R}^4$ that satisfies

$$y_i \approx a_0 + a_1 u_i + a_2 v_i + a_3 u_i v_i \quad \forall i \in \{1, \ldots, m\}$$

This can be viewed as a matrix equation

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{y} \approx \underbrace{\begin{bmatrix} 1 & u_1 & v_1 & u_1v_1 \\ 1 & u_2 & v_2 & u_2v_2 \\ 1 & u_3 & v_3 & u_3v_3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & u_m & v_m & u_mv_m \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix}}_{x}$$

This is in the form $y \approx Ax$. Assuming the matrix A has linearly independent columns, we know $A^\top A$ is invertible and we have solution

$$x^* = (a_0^*, a_1^*, a_2^*, a_3^*) = (A^\top A)^{-1} A^\top y$$

which is called the *least squares solution*, also called the *least squares fit*. Strictly speaking, this uses the sum of squared error as a minimization objective: It minimizes $\|y - Ax\|^2$ over all $x \in \mathbb{R}^4$. The A matrix is more likely to have linearly independent columns when it is tall, meaning the number of data points m is much larger than the number of columns. Furthermore, accuracy of the approximation tends to improve as we get more data. If A does *not* have linearly independent columns then the equation $A^\top y = A^\top Ax$ has infinitely many solutions $x \in \mathbb{R}^4$ and any solution can be used to produce the least squares fit.

For this problem, since $x^* = (a_0^*, a_1^*, a_2^*, a_3^*)$ satisfies $A^\top y = (A^\top A)x^*$ (which specifies four linear equations), it can be seen (from the first of the four equations) that the optimal coefficients a_i^* must satisfy

$$\frac{1}{m}\sum_{i=1}^{m} y_i = a_0^* + a_1^*\left(\frac{1}{m}\sum_{i=1}^{m} u_i\right) + a_2^*\left(\frac{1}{m}\sum_{i=1}^{m} v_i\right) + a_3^*\left(\frac{1}{m}\sum_{i=1}^{m} u_i v_i\right)$$

In particular, the least squares coefficients *exactly* fit the empirical averages of the data. This is a property of least squares that holds more broadly for other problems (see Exercises 9.32 and 9.34). It is also related to the *unbiased* property described in Section 9.13 for solutions to minimum mean square error estimation problems.

Why is it a good idea to find coefficients $x = (a_0, a_1, a_2, a_3)$ to minimize $\|y - Ax\|^2$? Intuitively, this can be understood as choosing a coefficient vector that *best explains the data*. To understand this, imagine that the true coefficients are exactly the (a_0, a_1, a_2, a_3) values. Then

$$y_i - (a_0 + a_1u_i + a_2v_i + a_3u_iv_i) = N_i \quad \forall i \in \{1, \ldots, m\}$$

where we have used (9.18) with the assumption $a_j = c_j$ for $j \in \{0, 1, 2, 3\}$. So

$$y - Ax = (N_1, \ldots, N_m) \quad , \quad \|y - Ax\|^2 = \sum_{i=1}^{m} N_i^2 \tag{9.19}$$

Then, the coefficients (a_0, a_1, a_2, a_3) that "best explain" the data can be viewed as the ones that make the corresponding noise sequence have a minimized sum of squares. This is a standard approach that works well in many contexts.

9.8.1. Gaussian noise

In the case when $\{N_i\}_{i=1}^m$ is an independent and identically distributed (i.i.d.) sequence of Gaussian random variables with mean 0 and unknown variance $\sigma^2 > 0$, the least-squares vector $x^* = (a_0^*, a_1^*, a_2^*, a_3^*)$ is the *max-likelihood estimate* of the true (c_0, c_1, c_2, c_3) values. This is because it maximizes the joint PDF of noise when this joint PDF is evaluated at vector $y - Ax$, which by (9.19) is the vector of noise values that match the data under the assumed coefficients x. Indeed, given y and A, we have

$$\begin{aligned} \arg\max_{x \in \mathbb{R}^4} \{f_{N_1,\ldots,N_m}(y - Ax)\} &= \arg\max_{x \in \mathbb{R}^4} \left\{ \frac{1}{(2\pi\sigma^2)^{m/2}} \exp\left(\frac{-\|y - Ax\|^2}{2\sigma^2} \right) \right\} \\ &= \arg\min_{x \in \mathbb{R}^4} \left\{ \|y - Ax\|^2 \right\} = x^* \end{aligned}$$

Notice that the variance parameter σ^2 does not need to be known. The technique works just as well when the Gaussian noise $\{N_i\}_{i=1}^m$ is i.i.d. with unknown mean $\mu \neq 0$ and unknown variance σ^2 (so $N_i \sim N(\mu, \sigma^2)$). This is because we can define $\tilde{N}_i = N_i - \mu$ and $\tilde{c}_0 = c_0 + \mu$, so the model (9.18) becomes

$$y_i = \tilde{c}_0 + c_1 u_i + c_2 v_i + c_3 u_i v_i + \tilde{N}_i$$

so the unknown coefficient c_0 is replaced by another unknown coefficient $\tilde{c}_0$. In other words, without loss of generality, the noise can be assumed to have zero mean. For example, a model that has $(c_0, c_1, c_2, c_3) = (1, -3, 0.3, 0.9)$, $N_i \sim N(3.2, \sigma^2)$ is equivalent to a model that has $(c_0, c_1, c_2, c_3) = (4.2, -3, 0.3, 0.9)$, $N_i \sim N(0, \sigma^2)$.

What if we do not know whether or not the noise is Gaussian? The good news is that the least squares solution can still be applied. This is because it is *data driven* and does not rely on a specific probability model. It is known to work well for Gaussian noise, but also works well in other practical scenarios. One reason is that the true noise is typically a sum of many independent random disturbances, so the central limit theorem suggests (under a suitable scaling) that the noise has a near-Gaussian distribution.

9.8.2. General noise

What if we know the noise sequence $\{N_i\}_{i=1}^{\infty}$ is i.i.d. with a known PDF $f_N(\cdot)$ that is not necessarily Gaussian? In that case, we could use a max-likelihood approach that chooses a coefficient vector $x^* = (a_0, a_1, a_2, a_3)$ to maximize the new joint PDF:

$$x^* = \arg\max_{x\in\mathbb{R}^4}\{f_{N_1,\ldots,N_m}(y - Ax)\} = \arg\max_{x\in\mathbb{R}^4}\left\{\prod_{i=1}^{m} f_N(y_i - (Ax)_i)\right\} \tag{9.20}$$

where $(Ax)_i$ denotes the ith coordinate of vector Ax, for $i \in \{1, \ldots, m\}$. Doing this can bring performance gains in comparison to the standard least squares. There are two main disadvantages of this approach:

1. It requires the noise PDF $f_N(\cdot)$ to be known.
2. It is not always easy to perform the maximization (9.20). Indeed, for general PDFs $f_N(\cdot)$, the above maximization problem can be nonconvex and difficult to solve. In contrast, the Gaussian distribution yields a closed form solution to (9.20) that does not depend on the variance parameter σ^2.

Because of these disadvantages, one may still choose to use the easy-to-calculate least squares solution $x^* = (A^\top A)^{-1}A^\top y$. When the noise is not Gaussian, the least squares solution is no longer guaranteed to be the max-likelihood solution. However, it may still work well in practice.

9.9. Weighted least squares

Consider the same problem $y \approx Ax$ of the previous section. Define the error vector

$$e = y - Ax$$

The standard least squares solution of the previous section seeks to minimize

$$||e||^2 = \sum_{i=1}^{m} e_i^2$$

which puts equal weight on each of the m data samples. In some cases we may believe that certain data points are more important than others, so we seek to minimize

$$\sum_{i=1}^{m} w_i^2 e_i^2 \tag{9.21}$$

where w_i^2 are given nonnegative weights.

More generally, given any $m \times m$ matrix W we can minimize

$$||We||^2 = e^\top W^\top We$$

This is the same as (9.21) when W is a pure diagonal matrix:

$$W = \begin{bmatrix} w_1 & 0 & 0 & \cdots & 0 \\ 0 & w_2 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & w_m \end{bmatrix} \tag{9.22}$$

To minimize $||We||^2$, multiply $y \approx Ax$ by W to obtain $Wy \approx WAx$. Now define

$$\tilde{y} = Wy \quad , \quad \tilde{A} = WA$$

Then

$$Wy \approx WAx \iff \tilde{y} \approx \tilde{A}x$$

and the latter is a least squares problem with solution

$$\boxed{x^* = (\tilde{A}^\top \tilde{A})^{-1} \tilde{A}^\top \tilde{y} = (A^\top W^\top WA)^{-1} A^\top W^\top Wy} \tag{9.23}$$

which assumes $\tilde{A}$ has linearly independent columns (so that $\tilde{A}^\top \tilde{A}$ is invertible). This x^* solves the problem of finding x to minimize $\left\|\tilde{y} - \tilde{A}x\right\|^2$, where

$$\left\|\tilde{y} - \tilde{A}x\right\|^2 = (\tilde{y} - \tilde{A}x)^\top (\tilde{y} - \tilde{A}x) = (y - Ax)^\top W^\top W(y - Ax) = ||We||^2$$

and $e = y - Ax$. The diagonal form of W in (9.22) ensures the solution x^* in (9.23) minimizes (9.21). See also Exercise 9.16. The solution (9.23) is more general than anticipated: It minimizes $||W(y - Ax)||^2$ for *any* real $m \times m$ matrix W (assuming WA has linearly independent columns). This is useful for Section 9.9.1.

If the $m \times m$ matrix W is invertible, then $\tilde{A}$ has linearly independent columns if and only if A has linearly independent columns (see Exercise 5.9). If $\tilde{A}$ does *not* have linearly independent columns, the weighted least squares problem $\tilde{y} \approx \tilde{A}x$ is still solved by finding any solution x to $\tilde{A}^\top \tilde{A}x = \tilde{A}^\top y$, which is guaranteed to have at least one solution.

The weighted least squares solution can often be interpreted as a solution under a modified inner product. Specifically, if W is invertible it can be shown that

$$\langle u, v \rangle = u^\top W^\top Wv \quad \forall u, v \in \mathbb{R}^m$$

is a valid inner product on $\mathbb{R}^m$. Then, the weighted least squares solution x^* in (9.23) corresponds to the minimizer of $\langle y - Ax, y - Ax \rangle$ over all x. Then, x^* satisfies the orthogonality equations with respect to this new inner product:

$$\left(\langle y - Ax^*, s \rangle = 0 \quad \forall s \in Col(A) \right) \iff \left((y - Ax^*)^\top W^\top W v_j = 0 \quad \text{for all columns } v_j \text{ of } A \right)$$

9.9.1. Jointly Gaussian noise

The weighted least squares solution has a surprising connection to estimation in the context of non-i.i.d. noise. Suppose we have a system with input vector $t \in \mathbb{R}^k$ (for some $k \in \mathbb{N}$) and real output y that can be repeatedly used, so

$$y_i = c_0 + \sum_{j=1}^{n} c_j g_j(t_i) + N_i \quad \forall i \in \{1, 2, 3, \ldots\}$$

where $n \in \mathbb{N}$; $c_0, c_1, \ldots, c_n$ are unknown real coefficients; $g_i : \mathbb{R}^k \to \mathbb{R}$ are known functions; N_i is unknown noise. For example, we might have $k = 1$, $n = 2$, $g_1(t) = t$, $g_2(t) = t^2$. We take m measurements and obtain data (t_i, y_i) for $i \in \{1, \ldots, m\}$. Then

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & g_1(t_1) & g_2(t_1) & \cdots & g_n(t_1) \\ 1 & g_1(t_2) & g_2(t_2) & \cdots & g_n(t_2) \\ \vdots & \vdots & \vdots & \vdots & \\ 1 & g_1(t_m) & g_2(t_m) & \cdots & g_n(t_m) \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_n \end{bmatrix}}_{x} + \underbrace{\begin{bmatrix} N_1 \\ N_2 \\ \vdots \\ N_m \end{bmatrix}}_{N}$$

Let $N = (N_1, \ldots, N_m)$ be the noise vector. Assume that N is *jointly Gaussian* with zero mean and a known invertible covariance matrix Λ, so

$$\mathbb{E}[N] = (0, 0, \ldots, 0) \quad , \quad \Lambda = \mathbb{E}\left[NN^\top\right]$$

[If the noise has known nonzero mean μ_i for $i \in \{1, \ldots, m\}$, just reduce to the zero mean case by defining output $\tilde{y}_i = y_i - \mu_i$.] In particular, the noise is not necessarily i.i.d., so the covariance between N_i and N_j is given by the (i, j) entry of matrix Λ. It can be shown that every invertible covariance matrix Λ is symmetric and positive definite. By eigenvalue and eigenvector properties of symmetric and positive definite matrices developed in Chapter 15, it turns out that Λ and its inverse have the following decomposition

$$\Lambda = S \Sigma S^\top \quad , \quad \Lambda^{-1} = S \Sigma^{-1} S^\top$$

where Σ is a diagonal matrix with positive diagonal entries, S is a square matrix that

satisfies $S^\top S = I$ (called a *unitary matrix*). Let $\Sigma^{1/2}$ and $\Sigma^{-1/2}$ be the diagonal matrices with all diagonal entries of Σ taken to the power $1/2$ and $-1/2$, respectively. For example, for $m = 3$ we have

$$\Sigma = \begin{bmatrix} d_{11} & 0 & 0 \\ 0 & d_{22} & 0 \\ 0 & 0 & d_{33} \end{bmatrix} , \Sigma^{1/2} = \begin{bmatrix} d_{11}^{1/2} & 0 & 0 \\ 0 & d_{22}^{1/2} & 0 \\ 0 & 0 & d_{33}^{1/2} \end{bmatrix} , \Sigma^{-1/2} = \begin{bmatrix} d_{11}^{-1/2} & 0 & 0 \\ 0 & d_{22}^{-1/2} & 0 \\ 0 & 0 & d_{33}^{-1/2} \end{bmatrix}$$

Then

$$\Lambda = S\Sigma^{1/2} \underbrace{S^\top S}_{I} \Sigma^{1/2} S^\top \quad , \quad \Lambda^{-1} = S\Sigma^{-1/2} \underbrace{S^\top S}_{I} \Sigma^{-1/2} S^\top$$

Define

$$W = S\Sigma^{-1/2}S^\top \tag{9.24}$$

The matrix W is an *invertible and symmetric square root* of matrix Λ^{-1}:

$$W^2 = W^\top W = \Lambda^{-1} \tag{9.25}$$

The joint Gaussian PDF $f_{N_1,\ldots,N_m}(z)$ for the zero mean case is (see also Chapter 11.12):

$$f_{N_1,\ldots,N_m}(z) = \frac{1}{(2\pi)^{m/2}\det(\Lambda)^{1/2}} \exp\left(-\frac{z^\top \Lambda^{-1} z}{2}\right) \quad \forall z \in \mathbb{R}^m$$

where $\det(\Lambda)$ is the determinant as defined in Chapter 11.

In this case, the max-likelihood estimator for the coefficient vector $x = (a_0, a_1, \ldots, a_n) \in \mathbb{R}^{n+1}$ is the vector x^* given by

$$\begin{aligned} x^* &= \arg\max_{x\in\mathbb{R}^{n+1}} \{f_{N_1,\ldots,N_m}(y - Ax)\} \\ &= \arg\max_{x\in\mathbb{R}^{n+1}} \left\{ \frac{1}{(2\pi)^{m/2}\det(\Lambda)^{1/2}} \exp\left(-\frac{(y-Ax)^\top \Lambda^{-1}(y-Ax)}{2}\right)\right\} \\ &= \arg\min_{x\in\mathbb{R}^{n+1}} \left\{(y-Ax)^\top \Lambda^{-1}(y-Ax)\right\} \\ &\overset{(a)}{=} \arg\min_{x\in\mathbb{R}^{n+1}} \left\{(y-Ax)^\top W^\top W(y-Ax)\right\} \\ &= \arg\min_{x\in\mathbb{R}^{n+1}} ||W(y-Ax)||^2 \end{aligned}$$

where (a) uses (9.25). Therefore, x^* is precisely the solution to the weighted least squares problem (9.23). In this case, the weight matrix W is the $m \times m$ matrix in (9.24).

For intuition, suppose the random noise variables $\{N_i\}_{i=1}^m$ are mutually independent

with zero mean and *different variances* $\sigma_1^2, \sigma_2^2, \ldots, \sigma_m^2$. Then $\mathbb{E}\left[N_i N_j\right] = 0$ if $i \neq j$, and $\mathbb{E}\left[N_i^2\right] = \sigma_i^2$. So

- Λ is a $m \times m$ diagonal matrix with diagonal entries σ_i^2.
- W is a $m \times m$ diagonal matrix with diagonal entries $1/\sigma_i$.

This makes sense: For each $i \in \{1, \ldots, m\}$, the weight matrix W prioritizes the data point (t_i, y_i) with weight $1/\sigma_i$. Data with smaller noise variances have higher weight. Define $e = y - Ax$. Then this approach chooses $x \in \mathbb{R}^3$ to minimize $\sum_{i=1}^m (1/\sigma_i^2) e_i^2$.

9.9.2. Fitting a model to an unknown system

The least squares technique has the following interpretation. Suppose we have m data points $(t_i, y_i) \in \mathbb{R}^2$ about an unknown system. We can use this data to build a model of the system of the form

$$y_i = a_0 + \sum_{j=1}^{n} a_j g_j(t_i) + N_i \quad \forall i \in \{1, \ldots, m\}$$

where $g_1(t), \ldots, g_n(t)$ are some collection of functions that we think will be a good fit for the actual system, and for coefficients $a_0, a_1, \ldots, a_n$ that we can tweak to make the fit as good as possible. It is useful to make the model as simple as possible, so often we use $n = 1$ and $g_1(t) = t$, or $n = 2$ and $g_1(t) = t, g_2(t) = t^2$. Such models can be used to *predict* the output y_i when a future input t_i is used.

For example, suppose $m = 10$ and we have the 10 data points (t_i, y_i) plotted in Fig. 9.3. It is reasonable to explain this data using a *linear fit*: Draw a line through the data that has

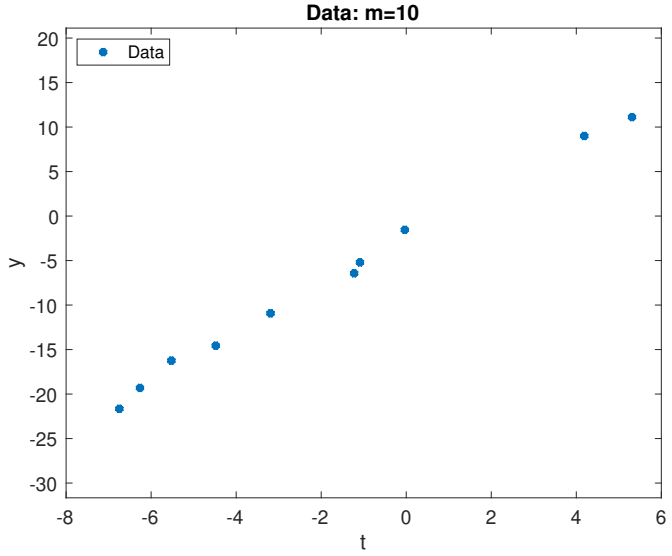


FIGURE 9.3. Ten data points (t_i, y_i). A linear model looks reasonable.

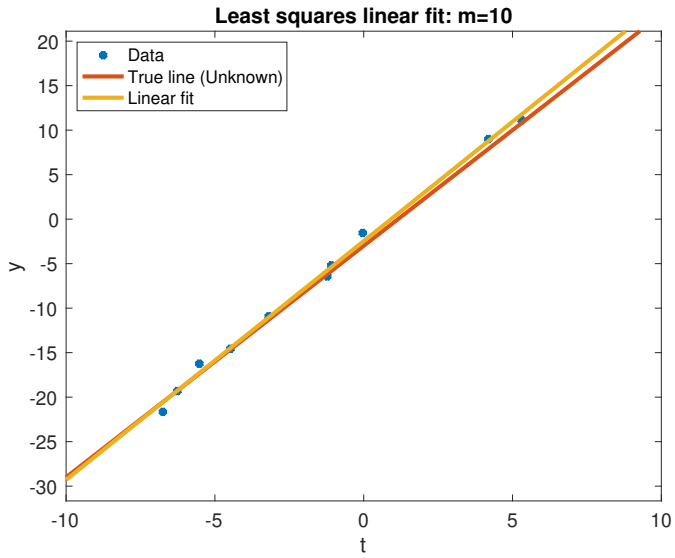


FIGURE 9.4. We have $m = 10$ data points. The true model was linear (red) with measurement noise. The least squares linear fit (gold) is a good fit.

the least sum squared error with the data points. So we build the following model:

$$y_i = a_0 + a_1 t_i + N_i \quad \forall i \in \{1, \ldots, m\}$$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{bmatrix}}_{y} \approx \underbrace{\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ 1 & t_3 \\ \vdots & \vdots \\ 1 & t_m \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a_0 \\ a_1 \end{bmatrix}}_{x}$$

Assume that A has linearly independent columns, which holds whenever there are indices i, j such that $t_i \neq t_j$. Then the coefficient vector $x = (a_0, a_1)$ that minimizes $\|y - Ax\|^2$ is

$$x^* = \begin{bmatrix} a_0^* \\ a_1^* \end{bmatrix} = (A^\top A)^{-1} A^\top y$$

This is a good engineering solution and produces a close fit to the original data, called a *least squares linear fit*. Fig. 9.4 plots the resulting linear fit as a gold line that roughly matches the data. In this particular example, the actual data *was* produced using a linear rule. Specifically, each t_i was chosen independently and uniformly over $[-10, 10]$. Then, we define $y_i = a_{0,true} + a_{1,true} t_i + N_i$ for some particular coefficients $a_{0,true}, a_{1,true}$ used to gen-

erate the red line in Fig. 9.4, and $\{N_i\}_{i=1}^m$ are independent and identically distributed (i.i.d.) Gaussian noise with distribution $N(0,1)$ (mean zero and variance 1). The least squares coefficients a_0^*, a_1^* are calculated only from the data points (t_i, y_i), without knowledge of this true model. They can also be used to accurately predict the output y when a new input t is used for the system: $y \approx a_0^* + a_1^* t$.

9.10. The dangers of overfitting

In general, we may tempted to explain data using a more complicated system model, such as a degree 9 polynomial model:

$$y_i = a_0 + a_1 t_i + a_2 t_i^2 + \ldots + a_9 t_i^9 + N_i \quad \forall i \in \{1, \ldots, m\} \tag{9.26}$$

One motivation for doing this is that, since we do not know the true system, the degree 9 polynomial model allows for a broader class potential systems. Indeed, this degree 9 polynomial allows the possibility that $a_2 = a_3 = \ldots = a_9 = 0$, so it includes a linear model as a special case. This system can be solved via least squares as follows:

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{10} \end{bmatrix}}_{y} \approx \underbrace{\begin{bmatrix} 1 & t_1 & t_1^2 & \cdots & t_1^9 \\ 1 & t_2 & t_2^2 & \cdots & t_2^9 \\ 1 & t_3 & t_3^2 & \cdots & t_3^9 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & t_{10} & t_{10}^2 & \cdots & t_{10}^9 \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_9 \end{bmatrix}}_{x}$$

so, assuming A has linearly independent columns, the coefficient vector is $x^* = (a_0^*, \ldots, a_9^*) = (A^\top A)^{-1} A^\top y$. We now show that using this complicated degree 9 structure can produce a more accurate fit to the data, but may not result in a better model of the system. This is related to the following lemma (see also Exercise 9.25).

LEMMA 9.2. *(Interpolating data points) Fix m as a positive integer. Let $t_1, \ldots, t_m$ be distinct real numbers. Let $y_1, \ldots, y_m$ be any (possibly repeated) real numbers. There is a polynomial function $p(t)$ of degree less than or equal to $m-1$ that exactly interpolates the points (t_i, y_i) for all $i \in \{1, \ldots, m\}$. That is*

$$p(t_i) = y_i \quad \forall i \in \{1, \ldots, m\}$$

PROOF. Define the following polynomial function:

$$p(t) = \sum_{i=1}^m y_i \left(\frac{\prod_{j \in \{1,\ldots,m\} \setminus \{i\}} (t - t_j)}{\prod_{j \in \{1,\ldots,m\} \setminus \{i\}} (t_i - t_j)} \right) \quad \forall t \in \mathbb{R}$$

This avoids division by zero because $t_i \neq t_j$ when $i \neq j$. Each term $i \in \{1, \ldots, m\}$ in the summation definition of $p(t)$ has degree either 0 (if $y_i = 0$) or $m-1$ (if $y_i \neq 0$). Thus, $p(t)$ has degree at most $m-1$. By construction we have $p(t_k) = y_k$ for all $k \in \{1, \ldots, m\}$. Indeed, if $k \in \{1, \ldots, m\}$, then we have for all $i \in \{1, \ldots, m\}$:

$$\frac{\prod_{j \in \{1,\ldots,m\} \setminus \{i\}} (t_k - t_j)}{\prod_{j \in \{1,\ldots,m\} \setminus \{i\}} (t_i - t_j)} = 0 \quad \text{if } i \neq k$$
$$\frac{\prod_{j \in \{1,\ldots,m\} \setminus \{i\}} (t_k - t_j)}{\prod_{j \in \{1,\ldots,m\} \setminus \{i\}} (t_i - t_j)} = 1 \quad \text{if } i = k$$

□

The above lemma ensures that for the 10 data points of Figs. 9.3 and 9.4, there is a degree 9 polynomial that exactly interpolates the data. Least squares finds an exact interpolation whenever one exists (zero error is certainly *least*). The degree 9 least squares fit is shown in Fig. 9.5. Unfortunately, the exact fit to the data produces an *unusually complicated* polynomial model, often called an *overfit*. The purple curve in Fig. 9.5 interpolates the data exactly, but is a wild degree 9 polynomial that wiggles randomly and attains points too large to be plotted in the figure. The simple linear fit (shown in gold) is a more accurate approximation of the true model (shown in red).

Fortunately, we can obtain accurate results even with a degree 9 polynomial model by

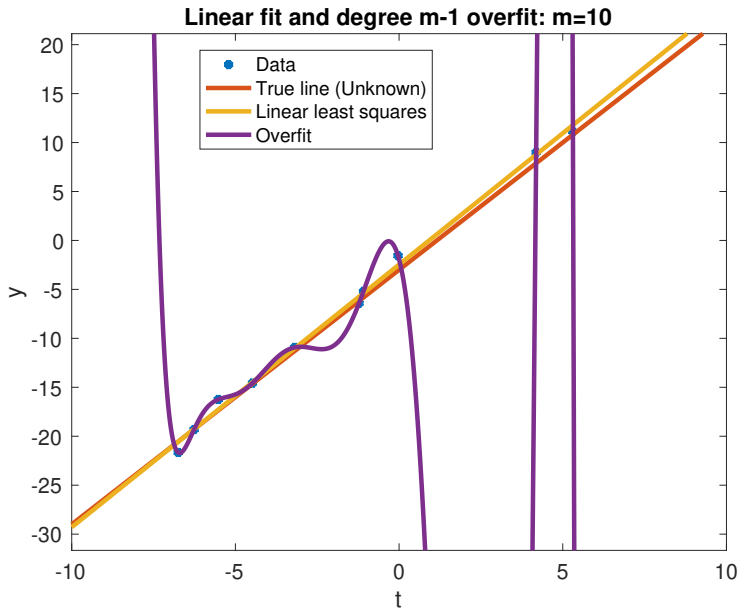


FIGURE 9.5. Purple curve is a least squares fit to polynomials of degree ≤ 9. It interpolates the data exactly, but is an *overfit* (overly complicated model).

using more data. Suppose we use $m > 10$ data points (t_i, y_i) to produce coefficients for a least squares fit to the degree 9 polynomial model:

$$y_i \approx \sum_{k=0}^{9} a_k t_i^k \quad \forall i \in \{1, \ldots, m\}$$

Fig. 9.6 shows the results for $m = 50$ (purple), together with the true line (red), and with the *linear* least squares fit (gold, calculated for $m = 50$ data points). The gold linear fit is almost an exact match to the true red line. The purple curve is the degree 9 polynomial, which now is a good fit in most of the interval $t \in [-10, 10]$ (although the error at the endpoints is apparent). Fig. 9.7 shows results for the case $m = 200$, where the degree 9 polynomial fit is almost the same as the linear fit. The 200 data points allow least squares to "learn" to make the coefficients $a_k \approx 0$ for the nonlinear terms $k \in \{2, \ldots, 9\}$.

Rules of thumb:

- Plotting the data (t_i, y_i) for $i \in \{1, \ldots, m\}$ helps build intuition for an accurate model.
- The number of data points should be significantly larger than the number of coefficients a_i in the model. Minimally, this means A should be a tall matrix. The taller the better: More data gives better results.

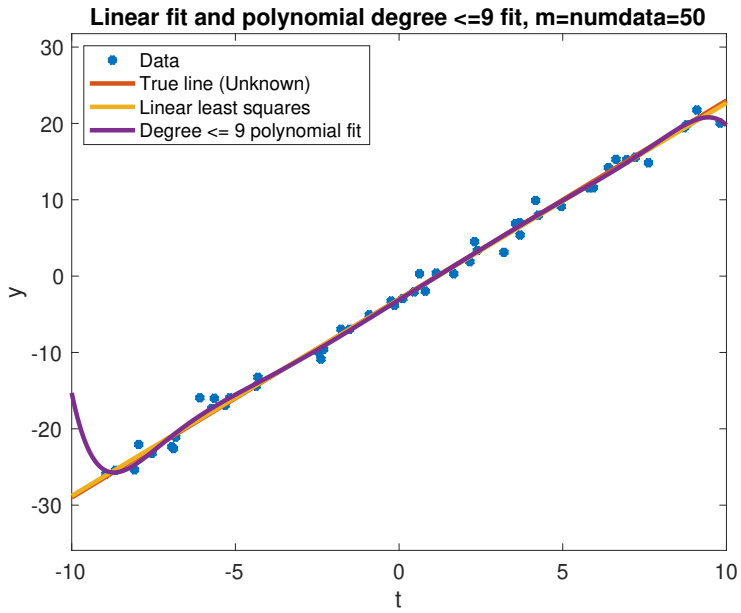


FIGURE 9.6. Now we have more data ($m = 50$). Least squares "learns" to fit the degree 9 polynomial to a *near-linear model in the interval of interest*.

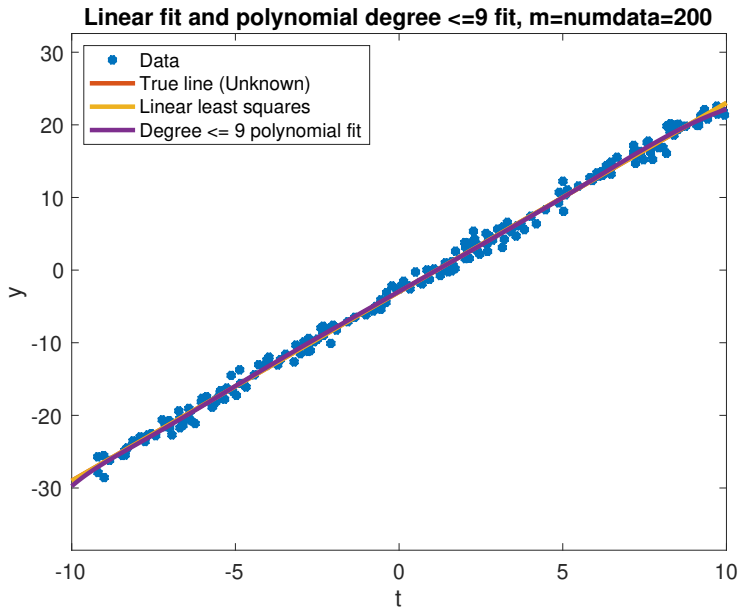


FIGURE 9.7. Now we have *much more data* ($m = 200$). Least squares fit to model $y \approx \sum_{k=0}^{9} \theta_k t^k$ "learns" to set $\theta_k \approx 0$ for $k \geq 2$.

9.11. Projecting with a general Gram matrix

Suppose V is a general real inner product space and S is a finite dimensional subspace (with dimension at least 1) such that

$$S = Span\{v_1, ..., v_k\}$$

for some vectors $v_1, ..., v_k$ that are not necessarily linearly independent. Then $1 \leq \dim(S) \leq k$. Fix $y \in V$. We want to project y onto S.

Let $p \in S$ be the projection (it exists and is unique). It must satisfy the orthogonality equations:

$$\langle y - p, s \rangle = 0 \quad \forall s \in S$$

Since $v_i \in S$ for all $i \in \{1, ..., k\}$ we obtain

$$\langle y - p, v_i \rangle = 0 \quad \forall i \in \{1, ..., k\}$$

That is, $\langle y, v_i \rangle = \langle p, v_i \rangle$ for all $i \in \{1, ..., k\}$. By symmetry of the inner product

$$\langle y, v_i \rangle = \langle v_i, p \rangle \quad \forall i \in \{1, ..., k\}$$

Since $p \in S$, we have $p = \sum_{j=1}^{k} \theta_j v_j$ for some scalars $\theta_j \in \mathbb{R}$ and so

$$\langle y, v_i \rangle = \left\langle v_i, \sum_{j=1}^{k} \theta_j v_j \right\rangle = \sum_{j=1}^{k} \theta_j \left\langle v_i, v_j \right\rangle \quad \forall i \in \{1, \ldots, k\}$$

These equations can be expressed in matrix form:

$$\underbrace{\begin{bmatrix} \langle y, v_1 \rangle \\ \langle y, v_2 \rangle \\ \vdots \\ \langle y, v_k \rangle \end{bmatrix}}_{h} = \underbrace{\begin{bmatrix} \langle v_1, v_1 \rangle & \langle v_1, v_2 \rangle & \cdots & \langle v_1, v_k \rangle \\ \langle v_2, v_1 \rangle & \langle v_2, v_2 \rangle & \cdots & \langle v_2, v_k \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_k, v_1 \rangle & \langle v_k, v_2 \rangle & \cdots & \langle v_k, v_k \rangle \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix}}_{\theta}$$

so $h = G\theta$, where G is the Gram matrix for $\{v_1, \ldots, v_k\}$ and h is the column vector of inner products between y and $\{v_1, \ldots, v_k\}$. Note that G depends only on $\{v_1, \ldots, v_k\}$, so it can be precomputed even before y is known. If we want to project *many* vectors $y \in V$ onto S, we can reuse the same G matrix (we only need to recompute vector h for each new y).

Existence of the projection ensures the equation $h = G\theta$ has at least one solution $\theta \in \mathbb{R}^k$. Every particular solution θ^* yields the same (unique) projection:

$$p = \sum_{j=1}^{k} \theta_j^* v_j$$

By Lemma 8.13, we know that if $\{v_1, \ldots, v_k\}$ are linearly independent then G is invertible, so the equation $h = G\theta$ has a unique solution $\theta^* = G^{-1}h$. If $S = Col(A)$, where A is a real $m \times k$ matrix with columns $\{v_1, \ldots, v_k\}$, the Gram matrix is $G = A^\top A$ and the solution $\theta^* = G^{-1}h$ reduces to the famous formula $\theta^* = (A^\top A)^{-1} A^\top y$ from Section 9.7.

If $\{v_1, \ldots, v_k\}$ is an *orthonormal basis for* S then $G = I$ and so $\theta^* = G^{-1}h = h$:

$$\theta_i^* = \langle y, v_i \rangle \quad \forall i \in \{1, \ldots, k\} \tag{9.27}$$

This is the same result obtained in the proof of the existence theorem (Theorem 9.2). This motivates Fourier series in the next section.

9.12. Fourier series

Fix $T > 0$. Let V be the real vector space of square integrable functions $f : [-T/2, T/2] \to \mathbb{R}$. Use the inner product:

$$\langle f, g \rangle = \int_{-T/2}^{T/2} f(t) g(t) dt \quad \forall f, g \in V$$

Use the standard equivalence concept $f \equiv g \iff \int_{-T/2}^{T/2}(f(t)-g(t))^2 dt = 0$. For every positive integer n it can be shown that the following $2n+1$ functions are orthonormal:

$$\left\{\frac{1}{\sqrt{T}}, \quad \left\{\sqrt{\frac{2}{T}}\cos\left(\frac{2\pi kt}{T}\right)\right\}_{k=1}^{n}, \quad \left\{\sqrt{\frac{2}{T}}\sin\left(\frac{2\pi kt}{T}\right)\right\}_{k=1}^{n}\right\} \tag{9.28}$$

This can be shown using the following three trig formulas that hold for all $x, y \in \mathbb{R}$:

$$\cos(x)\cos(y) = \frac{1}{2}(\cos(x-y)+\cos(x+y))$$
$$\cos(x)\sin(y) = \frac{1}{2}(-\sin(x-y)+\sin(x+y))$$
$$\sin(x)\sin(y) = \frac{1}{2}(\cos(x-y)-\cos(x+y))$$

So the Gram matrix for the $2n+1$ orthonormal functions in (9.28) is the $(2n+1)\times(2n+1)$ identity matrix I. Let S_n be the subspace of all linear combinations of the $2n+1$ functions in (9.28). Then any function $f(t) \in V$ has a projection onto S_n given by the *Fourier series*:

$$p_n(t) = a_0\frac{1}{\sqrt{T}} + \sum_{k=1}^{n} a_k\sqrt{\frac{2}{T}}\cos\left(\frac{2\pi kt}{T}\right) + \sum_{k=1}^{n} b_k\sqrt{\frac{2}{T}}\sin\left(\frac{2\pi kt}{T}\right) \quad \forall t \in \left[\frac{-T}{2}, \frac{T}{2}\right] \tag{9.29}$$

where, by (9.27), we have

$$a_0 = \left\langle f(t), \frac{1}{\sqrt{T}}\right\rangle \tag{9.30}$$

$$a_k = \left\langle f(t), \sqrt{\frac{2}{T}}\cos\left(\frac{2\pi kt}{T}\right)\right\rangle \quad \forall k \in \{1, \ldots, n\} \tag{9.31}$$

$$b_k = \left\langle f(t), \sqrt{\frac{2}{T}}\sin\left(\frac{2\pi kt}{T}\right)\right\rangle \quad \forall k \in \{1, \ldots, n\} \tag{9.32}$$

If $f(t)$ is an *even function* (so $f(-t) = f(t)$ for all t) it can be shown that $b_k = 0$ for all $k \in \{1, \ldots, n\}$; if $f(t)$ is an *odd function* (so $f(-t) = -f(t)$ for all t) it can be shown that $a_k = 0$ for all $k \in \{0, 1, \ldots, n\}$.

For example, consider $T = 1$ and the discontinuous even function $f \in V$ of Fig. 9.8. Its projection onto S_3, which uses only a linear combination of a constant function and the 3 lowest-frequency cosines, is also shown in Fig. 9.8. Its projection onto S_{10}, S_{50}, and S_{200} (again only using a constant and cosines) are shown in Figs. 9.9, 9.10, 9.11. Higher frequency cosines are needed to capture more detail in the original $f(t)$ function.

Since $f \in V$, it can be shown that the total energy in the error goes to zero as $n \to \infty$:

$$\lim_{n\to\infty} \|f - p_n\|^2 = \lim_{n\to\infty} \int_{-T/2}^{T/2} (f(t)-p_n(t))^2 dt = 0 \tag{9.33}$$

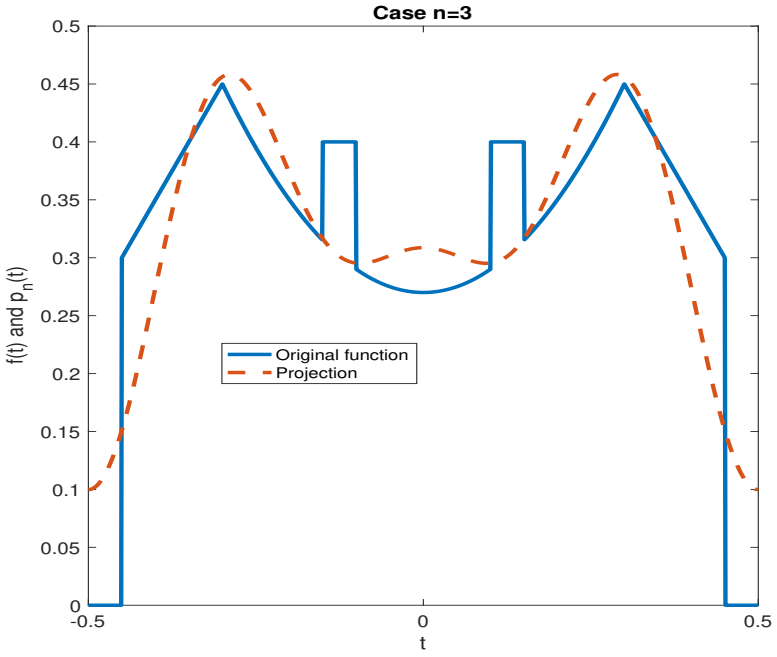


FIGURE 9.8. Solid: An even function $f(t)$ that resembles Hogwarts castle. Dashed: Projecting $f(t)$ onto S_3 (the case $n = 3$).

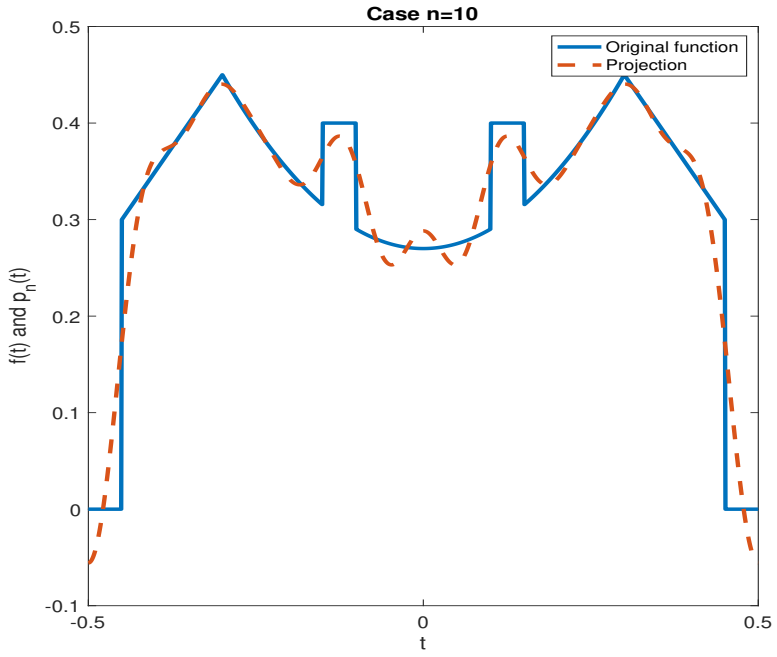


FIGURE 9.9. Projecting $f(t)$ onto S_{10} (the case $n = 10$).

The equality (9.33) is a version of the *Riesz-Fischer theorem*. The theorem ensures that if $f : [-T/2, T/2] \to \mathbb{R}$ is square integrable, and if p_n denotes the projection of f onto the subspace S_n for each $n \in \{1, 2, 3, \ldots\}$, then (9.33) holds. This demonstrates a remarkable property of sines and cosines at the integer frequencies $2\pi k/T$ for $k \in \{1, 2, 3, \ldots\}$. One related result is this: If $z : [-T/2, T/2] \to \mathbb{R}$ is a square integrable function that is orthogonal to all integer-frequency sine and cosine functions, and also orthogonal to the constant function 1, then $z \equiv 0$. The equation (9.33) can be viewed as a type of *convergence* of the p_n functions to the function f. The type of convergence in (9.33) is called L_2 *convergence* (it does not imply *pointwise convergence*). Figs. 9.10 and 9.11 show a well known *Gibbs overshoot* phenomenon at points of discontinuity of f. The width of the overshoot "wiggle" goes to zero as $n \to \infty$, as does its resulting squared integral contribution to the error. However, the height of the overshoot is persistent and is known to be about 9% of the discontinuous jump. Fig. 9.11 suggests the jump is less than 9% only because it samples the function at gridpoints along the horizontal axis that do not necessarily include the peak overshoot.

THEOREM 9.4. *(Parseval's identity) If $f \in V$ then $\|f\|^2 < \infty$, where $\|\cdot\|$ uses the standard inner product norm, and*

$$\|f\|^2 = \|f - p_n\|^2 + \|p_n\|^2 \qquad \forall n \in \mathbb{N}$$

where p_n is the orthogonal projection of f onto the subspace S_n spanned by the $2n + 1$ functions

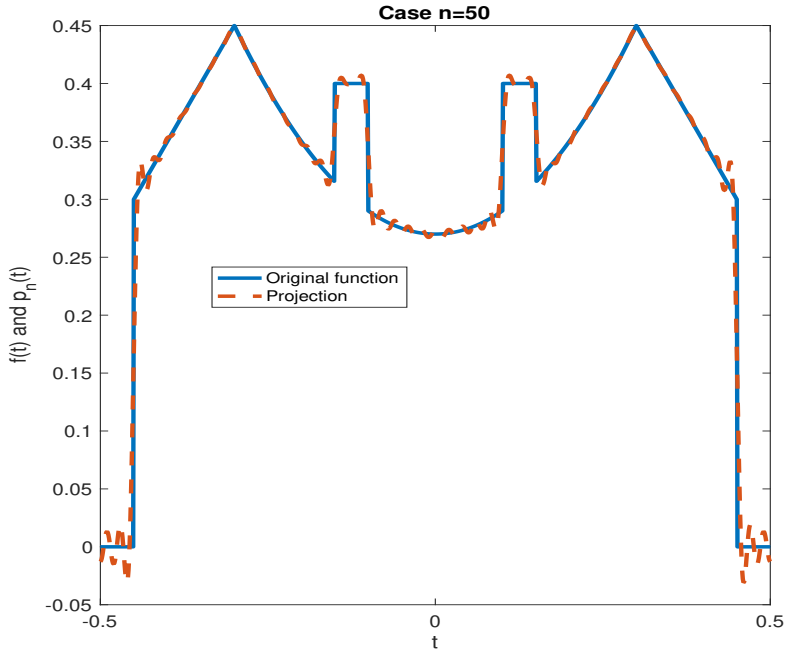


FIGURE 9.10. Projecting $f(t)$ onto S_{50} (the case $n = 50$).

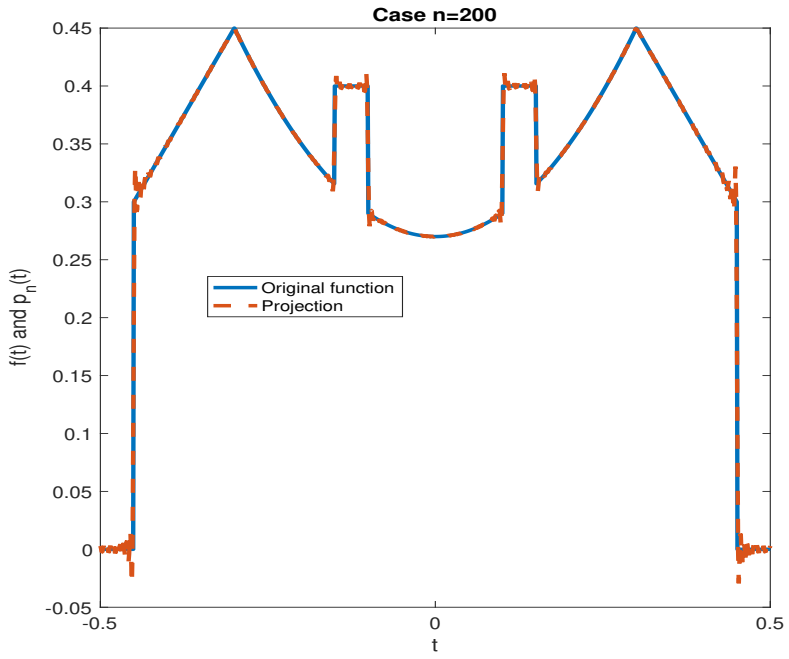


FIGURE 9.11. Projecting $f(t)$ onto S_{200} (the case $n = 200$).

in (9.28). *Also*

$$||f||^2 = \sum_{k=0}^{\infty} a_k^2 + \sum_{k=1}^{\infty} b_k^2$$

where a_k *and* b_k *are given in* (9.30)-(9.32).

PROOF. By definition of $f \in V$ we have $||f||^2 < \infty$. Then for all $n \in \mathbb{N}$

$$\begin{aligned} ||f||^2 &= ||f - p_n + p_n||^2 \\ &= ||f - p_n||^2 + ||p_n||^2 + 2\underbrace{\langle f - p_n, p_n \rangle}_{0} \\ &= ||f - p_n||^2 + ||p_n||^2 \end{aligned}$$

which uses the fact that p_n is the projection of f onto subspace S_n and so the orthogonality conditions ensure $f - p_n$ is orthogonal to everything in S_n, including $p_n \in S_n$. Taking $n \to \infty$ and using (9.33) gives

$$||f||^2 = \lim_{n\to\infty} ||p_n||^2$$

On the other hand, orthonormality of the component functions of p_n in (9.29) gives (by Lemma 8.6)

$$||p_n||^2 = \sum_{k=0}^{n} a_k^2 + \sum_{k=1}^{n} b_k^2 \quad \forall n \in \mathbb{N}$$

□

9.12.1. Any interval of size T

Fix $T > 0$ and $d \in \mathbb{R}$. The same treatment can be used for the real vector space V of square integrable functions $f : [d, d+T] \to \mathbb{R}$ with inner product

$$\langle f, g \rangle = \int_d^{d+T} f(t)g(t)dt \quad \forall f, g \in V$$

The same functions in (9.28) can be shown to be orthonormal over this new interval $[d, d+T]$. Then, the projection p_n of a function $f \in V$ onto the subspace S_n, the subspace spanned by the $2n+1$ orthonormal functions in (9.28), has the same form as (9.29), with the exception of being defined over $t \in [d, d+T]$ instead of $t \in [-T/2, T/2]$. The coefficients a_k and b_k still satisfy (9.30)-(9.32) with the exception that the inner product is now defined by integration over the interval $[d, d+T]$. While $p_n(t)$ is formally defined only over the interval $t \in [d, d+T]$, it repeats T-periodically when extended to $t \in \mathbb{R}$. This is because the $\sin(2\pi kt/T)$ and $\cos(2\pi kt/T)$ functions that are used are T-periodic for all $k \in \mathbb{N}$. If this extended $p_n(t)$ function is even, which can be assessed by its behavior over the interval $[-T/2, T/2]$, then it uses only $\cos(\cdot)$ functions.

9.13. Minimum mean square error estimation

A classic problem in probability theory is estimating an unknown random variable X given a collection of (known) observations $(Y_1, ..., Y_n)$. The assumption is that each random variable Y_i may have some dependency on X. Therefore, each Y_i can help to construct a good estimate for X. Suppose we form an *estimate* $\hat{X}$ as some function of the observations:

$$\hat{X} = f(Y_1, Y_2, ...Y_n)$$

where $f : \mathbb{R}^n \to \mathbb{R}$ maps the measurement vector to a real number.[2] There is no randomness in the function f. However, evaluating f at the random input vector $(Y_1, ..., Y_n)$ produces a random variable output $\hat{X}$. The goal is to find a function f that minimizes the *mean square error* (MSE):

$$MSE = \mathbb{E}\left[(X - \hat{X})^2\right]$$

To ensure MSE is well defined, it is assumed that $\mathbb{E}\left[X^2\right] < \infty$ and $\mathbb{E}\left[\hat{X}^2\right] < \infty$.

It is difficult to minimize MSE over all possible functions because the solution requires

[2] Formally, we should add the mild assumption that the function $f : \mathbb{R}^n \to \mathbb{R}$ is *Borel measurable*, so $f(Y_1, ..., Y_n)$ is a valid random variable. The set of Borel measurable functions contains all functions of practical interest, including all continuous functions (hence all linear and affine functions).

knowledge of the joint distribution for $(X, Y_1, ..., Y_n)$.[3] Instead, we consider minimizing MSE over the restricted set of all functions of the form

$$f(y_1, ..., y_n) = \theta_0 + \sum_{i=1}^{n} \theta_i y_i \quad \forall\, y \in \mathbb{R}^n$$

for some coefficients θ_i. Strictly speaking, such functions f are *affine* (linear plus a constant). However, this is called *linear estimation* because it considers linear combinations of $\{1, Y_1, ..., Y_n\}$. We seek $\theta = (\theta_0, \theta_1, \ldots, \theta_n)$ to solve

$$\text{Minimize:} \quad \mathbb{E}\left[(X - \hat{X})^2\right] \tag{9.34}$$

$$\text{Subject to:} \quad \hat{X} = \theta_0 + \sum_{i=1}^{n} \theta_i Y_i \tag{9.35}$$

$$(\theta_0, \theta_1, ..., \theta_n) \in \mathbb{R}^{n+1} \tag{9.36}$$

If $X, Y_1, ..., Y_n$ all have finite second moments, then $\mathbb{E}\left[(X - \hat{X})^2\right]$ is finite for all choices of $\theta_0, \theta_1, ..., \theta_n$.

The problem (9.34)-(9.36) can be equivalently posed as a *projection problem*: Let V be the vector space of random variables (on the probability space of interest) that have finite second moment. We assume $X, Y_1, ..., Y_n \in V$. Use the standard inner product $\langle X, Y \rangle = \mathbb{E}[XY]$ for all $X, Y \in V$, with the standard equivalence concept $X \equiv Y$ if and only if $P[X = Y] = 1$. Let $\hat{X}$ be the projection of the unknown random variable X onto $Span\{1, Y_1, ..., Y_n\}$. Then $\hat{X}$ is a linear combination of $\{1, Y_1, ..., Y_n\}$ that minimizes $\|X - \hat{X}\|^2$, which is equivalent to minimizing MSE because

$$\left\|X - \hat{X}\right\|^2 = \left\langle X - \hat{X}, X - \hat{X} \right\rangle = \mathbb{E}\left[(X - \hat{X})^2\right]$$

Our known solution from Section 9.11 is

$$\hat{X} = \theta_0 + \sum_{i=1}^{n} \theta_i Y_i$$

where $\theta = (\theta_0, \theta_1, \ldots, \theta_n)$ is a vector that solves the equation $h = G\theta$, where G is the Gram

[3] In probability theory, it is well known that the (possibly nonlinear) estimator that minimizes MSE is $\hat{X} = \mathbb{E}[X|Y_1, ..., Y_n]$, which, by definition of conditional expectation, is indeed a (Borel measurable) function of $(Y_1, ..., Y_n)$. The conditional expectation can be difficult to calculate and generally requires knowledge of the full joint distribution of $(X, Y_1, ..., Y_n)$. A standard result in probability theory is that, when $(X, Y_1, ..., Y_n)$ have a *jointly Gaussian* distribution, the optimal estimator has the linear form $\hat{X} = \theta_0 + \sum_{i=1}^{n} \theta_i Y_i$ with coefficients θ_i that solve (9.37).

matrix for $\{1, Y_1, \ldots, Y_n\}$ and h is the vector of inner products between X and $\{1, Y_1, \ldots, Y_n\}$:

$$\underbrace{\begin{bmatrix} \mathbb{E}[X] \\ \mathbb{E}[XY_1] \\ \mathbb{E}[XY_2] \\ \vdots \\ \mathbb{E}[XY_n] \end{bmatrix}}_{h} = \underbrace{\begin{bmatrix} 1 & \mathbb{E}[Y_1] & \mathbb{E}[Y_2] & \cdots & \mathbb{E}[Y_n] \\ \mathbb{E}[Y_1] & \mathbb{E}\left[Y_1^2\right] & \mathbb{E}[Y_1Y_2] & \cdots & \mathbb{E}[Y_1Y_n] \\ \mathbb{E}[Y_2] & \mathbb{E}[Y_2Y_1] & \mathbb{E}\left[Y_2^2\right] & \cdots & \mathbb{E}[Y_2Y_n] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[Y_n] & \mathbb{E}[Y_nY_1] & \mathbb{E}[Y_nY_2] & \cdots & \mathbb{E}\left[Y_n^2\right] \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_n \end{bmatrix}}_{\theta} \tag{9.37}$$

The solution depends only on the statistical information $\mathbb{E}[X]$, $\mathbb{E}[Y_i]$, $\mathbb{E}[XY_i]$ and $\mathbb{E}\left[Y_iY_j\right]$ for all $i, j \in \{1, \ldots, n\}$. Assuming G is invertible, the unique optimal coefficients are $\theta^* = G^{-1}h$.

The first row of the matrix equality (9.37) is:

$$\mathbb{E}[X] = \theta_0 + \sum_{i=1}^{n} \theta_i \mathbb{E}[Y_i]$$

which means that, if the optimal coefficients $\theta_0, \theta_1, \ldots, \theta_n$ are used, then the mean of X matches the mean of the optimal linear estimator $\hat{X} = \theta_0 + \sum_{i=1}^{n} \theta_i Y_i$. The optimal linear estimator is said to be *unbiased* because $\mathbb{E}[X] = \mathbb{E}\left[\hat{X}\right]$.

A classic scenario is when each Y_i is an attenuated version of X plus independent noise:

$$Y_i = \rho_i X + N_i \quad \forall i \in \{1, \ldots, n\} \tag{9.38}$$

where $\rho_1, \ldots, \rho_n$ are known real numbers; N_i is a noise variable with mean zero and finite variance σ_i^2 for each $i \in \{1, \ldots, n\}$; $\{X, N_1, \ldots, N_n\}$ are mutually independent. This scenario holds when different sensing devices are used to sense X, each device i having a different attenuation coefficient ρ_i and independent noise. Assume X has mean m_X and variance σ_X^2. The Gram matrix G and the corresponding h vector can be easily computed as functions of m_X, σ_X, ρ_i, σ_i^2. For example, (9.38) implies

$$\begin{aligned} \mathbb{E}[Y_i] &= \mathbb{E}[\rho_i X + N_i] = \rho_i \mathbb{E}[X] + \mathbb{E}[N_i] = \rho_i m_X \quad \forall i \in \{1, \ldots, n\} \\ \mathbb{E}[XY_i] &= \mathbb{E}\left[\rho_i X^2 + XN_i\right] = \rho_i \mathbb{E}\left[X^2\right] + \mathbb{E}[X]\mathbb{E}[N_i] = \rho_i(\sigma_X^2 + m_X^2) \quad \forall i \in \{1, \ldots, n\} \end{aligned}$$

where the last equality uses the fact that X and N_i are independent to ensure $\mathbb{E}[XN_i] = \mathbb{E}[X]\mathbb{E}[N_i] = 0$. Computation of $\mathbb{E}\left[Y_iY_j\right]$ for all $i, j \in \{1, \ldots, n\}$ is similar and is left as Exercise 9.52 (make sure to treat the cases $i = j$ and $i \neq j$ separately). Since $\{X, N_1, \ldots, N_n\}$ are mutually independent, hence pairwise uncorrelated, it can be shown that if each

noise variable N_i has a positive variance σ_i^2, then none of the random variables in the list $\{1, Y_1, ..., Y_n\}$ is equivalent to a linear combination of the others. This means $\{1, Y_1, ..., Y_n\}$ is linearly independent (as vectors in the vector space V) and the corresponding $(n+1)\times(n+1)$ Gram matrix G is invertible (see related Exercise 9.63).

While the specific model $Y_i = \rho_i X + N_i$ of the previous paragraph is useful, one reason that equation (9.37) is so powerful is that it holds for *any probability model* between the unknown X and the observation vector $(Y_1, ..., Y_n)$, provided that the random variables $X, Y_1, ..., Y_n$ have finite second moments. This includes *unknown probability models*. The statistical information $\mathbb{E}[X], \mathbb{E}[Y_i], \mathbb{E}[XY_i]$ and $\mathbb{E}\left[Y_iY_j\right]$ can be estimated by m independent measurements, called *ground-truth data* $(X[u], Y_1[u], \ldots, Y_n[u])$ for $u \in \{1, ..., m\}$:

$$\mathbb{E}[X] \approx \frac{1}{m}\sum_{u=1}^{m} X[u] \tag{9.39}$$

$$\mathbb{E}[Y_i] \approx \frac{1}{m}\sum_{u=1}^{m} Y_i[u] \tag{9.40}$$

$$\mathbb{E}[XY_i] \approx \frac{1}{m}\sum_{u=1}^{m} X[u]Y_i[u] \tag{9.41}$$

$$\mathbb{E}\left[Y_iY_j\right] \approx \frac{1}{m}\sum_{u=1}^{m} Y_i[u]Y_j[u] \tag{9.42}$$

which, by the law of large numbers (LLN), is accurate when the number of measurements m is large.

Remarkably, substituting the LLN estimates (9.39)-(9.42) into (9.37) is exactly the same as directly using the m data points

$$(X[u], Y_1[u], \ldots, Y_n[u]) \quad \forall u \in \{1, ..., m\}$$

in a linear least squares estimation to solve $X[u] \approx \theta_0 + \sum_{i=1}^{n} \theta_i Y_i[u]$ for $u \in \{1, ..., m\}$. Indeed, as described in Section 9.8, we can set up the least squares problem directly from the data:

$$\underbrace{\begin{bmatrix} X[1] \\ X[2] \\ \vdots \\ X[m] \end{bmatrix}}_{x} \approx \underbrace{\begin{bmatrix} 1 & Y_1[1] & Y_2[1] & \cdots & Y_n[1] \\ 1 & Y_1[2] & Y_2[2] & \cdots & Y_n[2] \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & Y_1[m] & Y_2[m] & \cdots & Y_n[m] \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_n \end{bmatrix}}_{\theta} \tag{9.43}$$

With A, θ, x as defined in the above underbraces, we have

$$A^\top A = m\tilde{G}$$
$$A^\top x = m\tilde{h}$$

where $\tilde{G}$ and $\tilde{h}$ are LLN approximations of G and h from (9.37) that are formed by substituting the LLN estimates (9.39)-(9.42). Thus

$$A^\top x = A^\top A\theta \iff \tilde{h} = \tilde{G}\theta \tag{9.44}$$

Finally, if A has linearly independent columns then both $A^\top A$ and $\tilde{G}$ are invertible and the θ vector from (9.44) satisfies

$$\theta = (A^\top A)^{-1}A^\top x = \tilde{G}^{-1}\tilde{h}$$

In other words, obtaining θ from the least squares problem (9.43) is the same as obtaining θ from the MSE problem (9.37) by approximating G and h with the LLN estimates $\tilde{G}$ and $\tilde{h}$.

9.14. Projecting onto $Null(A)$

LEMMA 9.3. *(Projecting w onto $Null(A)$) Let A be a real $m \times n$ matrix. Fix $w \in \mathbb{R}^n$. Then*

a) The following equation has at least one solution $r \in \mathbb{R}^m$:

$$AA^\top r = Aw \tag{9.45}$$

Further, if r_1 and r_2 are two solutions, then $A^\top r_1 = A^\top r_2$.

b) If r^* is any particular *solution to* (9.45) *then the following vector p^* is the unique projection of w onto $Null(A)$:*

$$p^* = w - A^\top r^*$$

c) If $Null(A^\top) = \{0\}$ (which requires A to be either square or wide) then $AA^\top$ is invertible and the equation of part (a) has one unique solution. In particular

$$r^* = (AA^\top)^{-1}Aw$$
$$p^* = [I - A^\top(AA^\top)^{-1}A]w \tag{9.46}$$

where I is the $n \times n$ identity matrix. The matrix $[I - A^\top(AA^\top)^{-1}A]$ is the projection matrix *associated with projecting a vector $w \in \mathbb{R}^n$ onto $Null(A)$.*

PROOF. To prove (a), note that since $Null(A)$ is finite dimensional, there is a unique projection of w onto $Null(A)$. Let $p \in Null(A)$ be the unique projection. The orthogonality

equations give

$$(w - p)^\top s = 0 \quad \forall s \in Null(A)$$

Thus, $w - p$ is orthogonal to everything in $Null(A)$, so

$$w - p \in Null(A)^\perp = Col(A^\top)$$

which means

$$w - p = A^\top r \tag{9.47}$$

for some vector $r \in \mathbb{R}^m$. Since $p \in Null(A)$, multiplying both sides of the above by A gives

$$Aw = AA^\top r$$

which proves existence of an r that solves (9.45). Further, if r_1 and r_2 are two solutions then $AA^\top(r_1 - r_2) = 0$ and so $r_1 - r_2 \in Null(AA^\top) = Null(A^\top)$ (recall Lemma 8.5), so $A^\top(r_1 - r_2) = 0$ and $A^\top r_1 = A^\top r_2$.

To prove (b), let r^* be a particular solution to (9.45). Define $p^* = w - A^\top r^*$. We know the projection p satisfies (9.47), so $p = w - A^\top r$. Then

$$p - p^* = (w - A^\top r) - (w - A^\top r^*) = A^\top(r^* - r)$$

However, both r and r^* are solutions to (9.45) and so the end of part (a) implies $A^\top r^* = A^\top r$. Then the right-hand-side of the above equality is 0, meaning $p = p^*$. So p^* is the unique projection. The proof of (c) is an exercise [use parts (a) and (b)]. □

The reader is invited to verify that the projection matrix $[I - A^\top(AA^\top)^{-1}A]$ satisfies

$$[I - A^\top(AA^\top)^{-1}A]w \in Null(A) \quad \forall w \in \mathbb{R}^n$$

$$[I - A^\top(AA^\top)^{-1}A]^2 = [I - A^\top(AA^\top)^{-1}A]$$

9.15. Projecting onto an affine set

Let V be a vector space over a general field $\mathbb{F}$. A set $A \subseteq V$ is *affine* if it is either empty or has the form

$$A = S + a_0 = \{s + a_0 \in V : s \in S\}$$

for some $a_0 \in V$ and some subspace $S \subseteq V$. (This uses the Minkowski set sum notation.) Details about affine sets are given in Appendix C. Some facts:

- An affine set A is a subspace if and only if $0 \in A$.

- If we are given a matrix $M \in \mathbb{F}^{m\times n}$ and a vector $b \in \mathbb{F}^m$, Theorem 4.1 ensures the set of all solutions to $Mx = b$ is an affine subset of $\mathbb{F}^n$.
- For the real field, it can be shown that a set $A \subseteq V$ is affine if and only if $ta + (1-t)b \in A$ for all $t \in \mathbb{R}$ and all $a, b \in A$. [A generalization of this fact is proven in Appendix C for any field $\mathbb{F}$ for which $1 + 1 \neq 0$.]

For real inner product spaces, projecting onto an affine set is similar to projecting onto a subspace. First consider projecting a point $y \in \mathbb{R}^m$ onto the affine set $v + Col(A)$, where v is a given vector in $\mathbb{R}^m$ and A is a given real $m \times n$ matrix. The projection, call it $g^* \in \mathbb{R}^m$, has the form $g^* = v + Ax^*$ for some optimally chosen $x^* \in \mathbb{R}^n$. So we solve:

$$\begin{aligned} &\text{Minimize:} && \|y - (v + Ax)\|^2 \\ &\text{Subject to:} && x \in \mathbb{R}^n \end{aligned}$$

Of course, the above problem is the same as

$$\begin{aligned} &\text{Minimize:} && \|(y - v) - Ax\|^2 \\ &\text{Subject to:} && x \in \mathbb{R}^n \end{aligned}$$

which is the problem of projecting $(y - v)$ onto $Col(A)$. Assuming A has linearly independent columns, the matrix $A^\top A$ is invertible and the optimally chosen x^* is

$$x^* = (A^\top A)^{-1} A^\top (y - v)$$

so the corresponding (unique) projection of y onto $v + Col(A)$ is $g^* = v + Ax^*$, that is

$$\boxed{g^* = [I - A(A^\top A)^{-1}A^\top]v + A(A^\top A)^{-1}A^\top y}$$

where I is the identity matrix with the same number of rows as A. If $v = 0$, this solution reduces to the standard projection of y onto $Col(A)$. Note that: (i) g^* is linear in the parameter vector (v, y); (ii) $A^\top g^* = A^\top y$.

Now consider the problem of projecting a point $w \in \mathbb{R}^n$ onto the affine set $\{x : Ax = b\}$, where A is a given $m \times n$ matrix and $b \in \mathbb{R}^m$. Assume there is at least one *particular solution* x_p to $Ax = b$ (so the set of interest is nonempty). Then, by Theorem 4.1,

$$\{x : Ax = b\} = x_p + Null(A) \tag{9.48}$$

where the Minkowski set sum is again used. Thus, the set of solutions to $Ax = b$ is indeed

an affine set. Our problem seeks an $x \in \mathbb{R}^n$ to solve

$$\text{Minimize:} \quad \|w - x\|^2 \tag{9.49}$$

$$\text{Subject to:} \quad Ax = b, x \in \mathbb{R}^n \tag{9.50}$$

By (9.48), we know our solution x^* has the form

$$x^* = x_p + z^*$$

for some $z^* \in Null(A)$. Thus, the problem can be expressed as finding $z \in \mathbb{R}^n$ to solve

$$\begin{aligned} \text{Minimize:} \quad & \left\|w - (x_p + z)\right\|^2 \\ \text{Subject to:} \quad & Az = 0,\ z \in \mathbb{R}^n \end{aligned}$$

which is the same as finding a $z^* \in Null(A)$ that is the projection of $(w - x_p)$ onto $Null(A)$. From the theory of projecting onto the null space given in Section 9.14, we know the following: If $Null(A^\top) = \{0\}$ (which holds when A has linearly independent rows), the unique solution z^* is

$$z^* = [I - A^\top (AA^\top)^{-1} A](w - x_p)$$

and so

$$\begin{aligned} x^* &= x_p + z^* \\ &= x_p + [I - A^\top (AA^\top)^{-1} A](w - x_p) \\ &= w - A^\top (AA^\top)^{-1} Aw + A^\top (AA^\top)^{-1} Ax_p \\ &= [I - A^\top (AA^\top)^{-1} A]w + A^\top (AA^\top)^{-1} b \end{aligned}$$

which uses the fact $Ax_p = b$.

In summary, if A has linearly independent rows the unique solution to (9.49)-(9.50) is

$$\boxed{x^* = [I - A^\top (AA^\top)^{-1} A]w + A^\top (AA^\top)^{-1} b} \tag{9.51}$$

One can directly verify (9.51) gives $Ax^* = b$ (regardless of w). The problem (9.49)-(9.50) has the meaningful interpretation of finding a solution to $Ax = b$ that is *closest to a given point* $w \in \mathbb{R}^n$. It is interesting that the solution x^* in (9.51) is linear in the parameter vector (w, b). If $w = 0$ then we seek a solution to $Ax = b$ that is *closest to the origin* and the solution (9.51) reduces to

$$x^* = A^\top (AA^\top)^{-1} b$$

which coincides with the particular solution suggested in Lemma 7.1a.

9.16. The projection operator

9.16.1. Properties of the closest point function

Let V be a real inner product space (possibly infinite dimensional). Let $\langle x, y \rangle$ denote the inner product. Let S be a (possibly infinite dimensional) subspace of V. Suppose that for each $y \in V$, there is a closest point in S to y, where distance is defined by the inner product norm. Theorem 9.1 ensures the closest point is unique. Define $P : V \to V$ as the function that maps each $y \in V$ to its unique closest point in S. It can be shown that:

1. $P : V \to V$ is a linear operator. [Prove this via the orthogonality conditions.]
2. $P(s) = s$ for all $s \in S$. [Clearly s is the closest point in S to itself.]
3. $S = Span(P)$. [A consequence of bullet 2 and the fact $P(y) \in S \quad \forall y \in V$.]
4. $P(P(y)) = P(y)$ for all $y \in V$. [A consequence of bullet 2 and the fact $P(y) \in S \quad \forall y \in V$.]
5. $\langle P(x), y \rangle = \langle x, P(y) \rangle$ for all $x, y \in V$. [See Exercise 9.58.]

The function $P : V \to V$ is called the *projection operator* associated with projecting vectors in V onto the given subspace S and using the given inner product. (Recall that an "operator" is just a linear transformation from a vector space to itself.)

9.16.2. A generalized theory of projections

DEFINITION 9.1. *Let V be a vector space over some field $\mathbb{F}$. A function $P : V \to V$ is a called a* projection operator *if it is a linear transformation that satisfies $P(P(y)) = P(y)$ for all $y \in V$.*

DEFINITION 9.2. *Let V be a real inner product space. A function $P : V \to V$ is called an* orthogonal projection operator *if it is a projection operator that satisfies*

$$\langle P(x), y \rangle = \langle x, P(y) \rangle \quad \forall x, y \in V$$

This definition of *projection operator* is general and does not require existence of an inner product. From bullets 1, 4, and 5 in the previous subsection, the closest point function defined there is both a projection operator and an orthogonal projection operator. We now show that when V is a vector space over field $\mathbb{R}$: (i) Every orthogonal projection operator $P : V \to V$ is a function that maps each $y \in V$ to its closest point in the subspace $Span(P)$, where distances are defined by the inner product norm; (ii) Every projection operator is an orthogonal projection operator with respect to some inner product.

LEMMA 9.4. *Let $P : V \to V$ be a projection operator. Define $S = Span(P)$. Then $P(s) = s \quad \forall s \in S$.*

PROOF. See Exercise 9.60. □

LEMMA 9.5. *Suppose V is a real inner product space. Let $P : V \to V$ be an orthogonal projection operator. Define $S = Span(P)$. Then for each $y \in V$, the vector $P(y)$ is the unique closest point in S to y (using the inner product norm to define distances).*

PROOF. Fix $y \in V$. By Theorem 9.3, it suffices to prove the orthogonality condition:

$$\langle y - P(y), s\rangle = 0 \quad \forall s \in S$$

For this, fix $s \in S$. We have

$$\begin{aligned}\langle y - P(y), s\rangle &= \langle y, s\rangle - \langle P(y), s\rangle \\ &\overset{(a)}{=} \langle y, s\rangle - \langle y, P(s)\rangle \\ &\overset{(b)}{=} \langle y, s\rangle - \langle y, s\rangle = 0\end{aligned}$$

where (a) uses the fact that P is an orthogonal projection operator; (b) uses Lemma 9.4. □

LEMMA 9.6. *Suppose V is a finite dimensional real vector space. Let $P : V \to V$ be a projection operator. Then there is an inner product on V under which P is an orthogonal projection operator.*

PROOF. Define $S = Span(P)$. We show there is an inner product under which P is the function that maps each $y \in V$ to its closest point in S. The result is trivial if $\dim(S) = 0$. Suppose $\dim(S)$ is a positive integer. Let $\{v_1, ..., v_k\}$ be a basis for S. Let $\{y_1, ..., y_m\}$ be additional vectors such that $\{v_1, ..., v_k, y_1, ..., y_m\}$ is a basis for V. Exercise 9.59 shows that $\{v_1, ..., v_k, y_1 - P(y_1), ..., y_m - P(y_m)\}$ is linearly independent. Then, Corollary 8.2 ensures there exists an inner product on V under which $\{v_1, ..., v_k, y_1 - P(y_1), ..., y_m - P(y_m)\}$ are orthonormal. Thus, for each $i \in \{1, ..., m\}$, we know $y_i - P(y_i)$ is orthogonal to all basis vectors of S, so it is orthogonal to all $s \in S$. So $P(y_i)$ satisfies the orthogonality conditions of Theorem 9.3, meaning that $P(y_i)$ is the unique closest point in S to y_i.

Also, since P is a projection operator, for all $j \in \{1, ..., k\}$ Lemma 9.4 ensures $P(v_j) = v_j$. Therefore, P maps each basis vector in $\{v_1, ..., v_k, y_1, ..., y_m\}$ to its unique closest point in S. By linearity of the closest point function, it follows that for each $y \in V$, $P(y)$ is the unique closest point in S to y. Then, bullet 5 of Section 9.16.1 ensures P is an orthogonal projection operator. □

The result of Lemma 9.6 can be extended to the case when V is an infinite dimensional real vector space using the concept of Hamel basis (discussed in Chapter 8.10.2).

9.17. Exercises

EXERCISE 9.1. *(Basic projections) Using the standard Euclidean inner product, project* $(1, 4, 5) \in \mathbb{R}^3$ *onto:*

a) $Span\{(1,-1,2)\}$

b) $Span\{(1,-1,2),(1,1,1)\}$.

EXERCISE 9.2. *(Multiple choice for projections) Let* V *be a real inner product space. Let* S *be a subspace of* V*. Fix* $y \in V$*. Suppose there exists a projection point* $p = Proj_S(y)$*. Then*

a) $p \in S$

b) $||y - p|| \leq ||y - s|| \quad \forall s \in S$

c) $\langle y - p, s \rangle = 0 \quad \forall s \in S$

d) p is the unique closest point in S to y

e) All of the above.

EXERCISE 9.3. *(Multiple choice for existence of projection) Let* V *be a real inner product space. Let* S *be a subspace of* V*. Fix* $y \in V$*. Which of the following is the* least restrictive *assumption that ensures a projection* $p = Proj_S(y)$ *exists? Note that* least restrictive *means* most general.

a) $\dim(V) < \infty$

b) $\dim(S) < \infty$

c) $V = \mathbb{R}^n$ *for some* $n \in \mathbb{N}$

d) $S = Span\{v_1, v_2, v_3\}$

e) $S = \{0\}$

EXERCISE 9.4. *(Multiple choice for square integrable functions) Let* V *be the real inner product space of square integrable functions* $f : [-1/2, 1/2] \to \mathbb{R}$*. Let* S *be a subspace of* V*. Fix* $y \in V$*. Fix* n *as an integer larger than 3. Which is always true?*

a) There exists a projection $p = Proj_S(y)$.

b) $\{1, \sqrt{2}\cos(2\pi t), \sqrt{2}\cos(4\pi t), \ldots, \sqrt{2}\cos(2\pi n t)\}$ *are orthonormal vectors in* V.

c) The function $f : [-1/2, 1/2] \to \mathbb{R}$ *is in* V*, where*

$$f(t) = \begin{cases} 0 & \text{if } t \leq 0 \\ 1/t & \text{if } 0 < t \leq 1 \end{cases}$$

d) S contains only continuous functions

e) S contains only bounded functions

EXERCISE 9.5. *(Projection operation is linear) Let* V *be a real inner product space. Let* S *be a finite dimensional subspace of* V*. Define the function* $T : V \to V$ *by*

$$T(y) = Proj_S(y) \quad \forall y \in V$$

a) Suppose $y \in S$*. What is* $T(y)$*?*

b) Suppose $y \in V$*. What is* $T(T(y))$*? What is* $T(T(T(y)))$*?*

c) Show that T is a linear function. Hint: Use the orthogonality conditions from Theorem 9.3.

EXERCISE 9.6. *(Project onto a line) Project* $(8,-1,3)$ *onto* $Span\{(1,1,1)\}$. *Find orthogonal vectors* w, p *such that* $p \in Span\{(1,1,1)\}$ *and* $(8,-1,3) = w + p$.

EXERCISE 9.7. *(Project onto another line) Project* $(4,-4,1,1)$ *onto* $Span\{(1,1,1,2)\}$. *Find orthogonal vectors* w, p *such that* $p \in Span\{(1,1,1,2)\}$ *and* $(4,-4,1,1) = w + p$.

EXERCISE 9.8. *(Projection onto a plane) Project* $(8,-1,3)$ *onto* $Span\{(1,1,1),(1,0,-1)\}$. *Find orthogonal vectors* w, p *such that* $p \in Span\{(1,1,1),(1,0,-1)\}$ *and* $(8,-1,3) = w + p$.

EXERCISE 9.9. *(Projecting onto a volume) Project* $(8,-1,3)$ *onto* $Span\{(1,1,1),(1,0,-1),(2,2,5)\}$. *Find orthogonal vectors* w, p *such that* $p \in Span\{(1,1,1),(1,0,-1),(2,2,5)\}$ *and* $(8,-1,3) = p + w$.

EXERCISE 9.10. *(Least squares with invertible* $A^\top A$*) Consider* y *and* A *defined below*

$$A = \begin{bmatrix} 5 & 1 & -1 \\ 0 & 3 & 0 \\ 2 & 2 & 2 \\ 4 & 18 & -1 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$$

a) Show A *has linearly independent columns. Find a vector* $x^* \in \mathbb{R}^3$ *that solves the least squares problem* $y \approx Ax$.

b) Compare y *with* Ax^*. *Is* $y \in Col(A)$? *Show that* $y - Ax^*$ *is orthogonal to all columns of* A.

c) Find $||y - Ax^*||^2$. *Test two other vectors* $x \in \mathbb{R}^3$ *and compare their* $||y - Ax||^2$ *values to* $||y - Ax^*||^2$.

EXERCISE 9.11. *(Least squares with noninvertible* $A^\top A$*) Consider* y *and* A *defined below*

$$A = \begin{bmatrix} 6 & 2 & 2 & 0 & 0 \\ 3 & 1 & 1 & 3 & 3 \\ 8 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \end{bmatrix}, \quad y = \begin{bmatrix} 5 \\ 5 \\ 5 \\ 5 \end{bmatrix}$$

a) Show $rank(A) = 3$. *Find a vector* $x^* \in \mathbb{R}^5$ *that solves the least squares problem* $y \approx Ax$.

b) Compare y *with* Ax^*. *Show that* $y - Ax^*$ *is orthogonal to all columns of* A.

c) Find a different vector $\tilde{x} \in \mathbb{R}^5$ *that solves the least squares problem* $y \approx Ax$. *Compare* $A\tilde{x}$ *and* Ax^*.

d) Is $y \in Col(A)$?

EXERCISE 9.12. *Design and solve your own least squares* $Ax \approx y$ *problem.*

EXERCISE 9.13. *Let V be the vector space of square integrable functions $f : [0,1] \to \mathbb{R}$. The inner product is $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$ for all $f, g \in V$. Given $f \in V$, we want to find a function $g \in V$ that satisfies:*

- $g(t) \in Span\{1, t, t^2\}$
- $g(0) = f(0)$
- $g(t) \approx f(t)$ *for all* $t \in [0,1]$

For this, we define $g(t) = f(0) + p(t)$ where $p(t)$ is the projection of $f(t) - f(0)$ onto $Span\{t, t^2\}$.

a) Show how to find optimal coefficients a_1, a_2 for $p(t) = a_1 t + a_2 t^2$.

b) Find a_1, a_2 for the case when $f(t) = e^t$. Plot to compare $f(t)$ and $g(t)$.

c) Find a_1, a_2 for the case when $f(t) = \cos(t)$. Plot to compare $f(t)$ and $g(t)$.

EXERCISE 9.14. *(Projecting functions) Let V be the real inner product space of square integrable functions $f : [0,1] \to \mathbb{R}$. Let $\langle f, g\rangle = \int_0^1 f(t)g(t)dt$. Project t^5 onto $Span\{t, t^3\}$ (treating t, t^3, t^5 as functions in V).*

EXERCISE 9.15. *Let V be a real inner product space. Let S be a finite dimensional subspace of V. Show that every vector $y \in V$ can be uniquely written $y = p + z$ where $p \in S$ and $z \in S^\perp$. Hint: Project y onto S.*

EXERCISE 9.16. *(Improving accuracy of $Ax \approx y$ in some components) Consider A, y in* (9.17).

a) Explain why Ax has a zero in row 2 for every $x \in \mathbb{R}^2$.

b) Find $v \in \mathbb{R}^2$ so that Av matches y perfectly in rows 1 and 3. Compare $\|Av - y\|^2$ to $\|Ax^* - y\|^2$ where x^* is the least squares solution given just below equation (9.17).

c) Use weighted least squares to find $z \in \mathbb{R}^2$ that minimizes $4(y_1 - p_1)^2 + (y_2 - p_2)^2 + 4(y_3 - p_3)^2 + (y_4 - p_4)^2$, where $p = Az = (p_1, p_2, p_3, p_4)$. Compare to parts (a) and (b).

EXERCISE 9.17. *(Function approximation) We want to approximate $\cos(t)$ over the interval $[0, \pi/2]$ using polynomials of degree less than or equal to 2. Define*

$$V = \{\text{square integrable functions } f : [0, \pi/2] \to \mathbb{R}\}$$

Define the subspace $S \subseteq V$ by

$$S = \{\text{polynomial functions } f : [0, \pi/2] \to \mathbb{R} \text{ of degree} \le 2\} = Span\{1, t, t^2\}$$

a) Find a function $p(t) \in S$ that minimizes $\int_0^{\pi/2} (\cos(t) - p(t))^2 dt$.

b) Use MATLAB *(or some other computing software) to plot the original function $\cos(t)$ over the interval $[0, \pi/2]$, together with your function $p(t)$ from part (a), and with the Taylor expansion $\cos(t) \approx 1 + 0t + \frac{-1}{2}t^2$.*

c) Explain why approximating using polynomial functions of degree less than or equal to 3

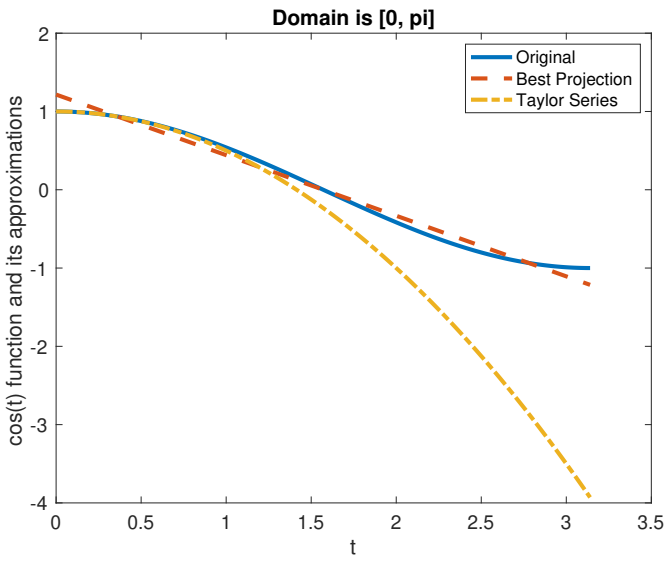


FIGURE 9.12. For Exercise 9.19: Approximating $\cos(t)$ over $t \in [0, \pi]$ by the projection $p(t) \in Span\{1, t, t^2\}$ and the Taylor expansion $\cos(t) \approx 1 - t^2/2 \in Span\{1, t, t^2\}$.

would give accuracy at least as good or better than the case of polynomial functions of degree less than or equal to 2.

EXERCISE 9.18. *(Function approximation 2) Repeat the previous problem 9.17, but approximating* $\cos(t)$ *over the interval* $[0, \pi]$ *using polynomials of degree less than or equal to 2. Find the function* $p(t)$ *and compare* $\cos(t)$, $p(t)$, *and the Taylor expansion for* $t \in [0, \pi]$.

EXERCISE 9.19. *(Function approximation 3) In the previous problems we approximated* $\cos(t)$ *over the interval* $[0, \pi]$ *using polynomials of degree less than or equal to 2. We considered projections* $p(t)$ *onto* $Span\{1, t, t^2\}$, *so* $p(t) = a + bt + ct^2$. *Due to the change in curvature of* $\cos(t)$ *over* $t \in [0, \pi]$, *the best projection is linear (so* $c^* = 0$*). Fig. 9.12 shows the* $\cos(t)$ *function, the projection* $p(t) = a^* + b^*t + c^*t^2$, *and the Taylor expansion* $\cos(t) \approx 1 - t^2/2$. *While the projection* $p(t)$ *fits the cosine function over the entire interval better than the Taylor expansion, it still has significant error. Indeed, a linear approximation to a cosine function is inherently inaccurate! To get a better approximation, we need to approximate the true curvature of* $\cos(t)$ *by adding a* t^3 *term. This term* t^3 *itself changes its curvature, so hopefully it can help to approximate the change in curvature of* $\cos(t)$ *over the interval* $[0, \pi]$.

The Taylor expansion does not change when allowing t^3 *terms: The Taylor expansion remains* $\cos(t) \approx 1 - t^2/2$. *However, the projection* $p(t)$ *will change in comparison to before. This problem is designed to see how much better the projection* $p(t) \in Span\{1, t, t^2, t^3\}$ *will do in comparison to using* $p(t) \in Span\{1, t, t^2\}$.

Formally, define V *as the real vector space of square integrable functions* $f : [0, \pi] \to \mathbb{R}$ *with inner product* $\langle f, g \rangle = \int_0^\pi f(t)g(t)dt$ *for all* $f, g \in V$. *Define subspace* $S = Span\{1, t, t^2, t^3\}$.

Consider the function $\cos(t) \in V$. *We want the projection of* $\cos(t)$ *onto the subspace S. The projection function* $p(t)$ *has the form* $p(t) = a + bt + ct^2 + dt^3$.

a) Curvature change: Show that $\frac{d^2}{dt^2}\cos(t) \leq 0$ *if* $t \in [0, \pi/2]$, *and greater than or equal to zero if* $t \in [\pi/2, \pi]$.

b) Curvature change: Show that $\frac{d^2}{dt^2}(t - \pi/2)^3 \leq 0$ *if* $t \in [0, \pi/2]$ *and greater than or equal to zero if* $t \in [\pi/2, \pi]$.

c) Show that $(t - \pi/2)^3 \in Span\{1, t, t^2, t^3\}$.

d) Use the orthogonality equations to solve for the best a^*, b^*, c^*, d^*.

e) On a graph, plot $\cos(t)$, *your projection* $p(t)$, *and the Taylor expansion* $1 - t^2/2$ *over* $t \in [0, \pi]$. *Discuss why the resulting accuracy for the new* $p(t)$ *function is much better in comparison to Fig. 9.12.*

EXERCISE 9.20. *(Orthogonal functions) Let V be the real inner product space of square integrable functions* $f : [0, 1] \to \mathbb{R}$ *with inner product* $\langle f, g \rangle = \int_0^1 f(t)g(t)dt$. *Find a nonzero function* $f \in V$ *that is orthogonal to both* t^2 *and* e^{-t}.

EXERCISE 9.21. *(More orthogonal functions) Let V be the space of square integrable functions* $f : [0, 2\pi] \to \mathbb{R}$ *with real inner product* $\langle f, g \rangle = \int_0^{2\pi} f(t)g(t)dt$. *Find a nonzero function* $f \in V$ *that is orthogonal to both* $\cos(t)$ *and* e^t.

EXERCISE 9.22. *Let V be a real inner product space. We have vectors* $\{v_1, ..., v_k\} \in V$. *Suppose* $y \in V$ *and* $\langle y, v_i \rangle = 0$ *for all* $i \in \{1, \ldots, k\}$. *Show that* $y \in (Span\{v_1, ..., v_k\})^{\perp}$.

EXERCISE 9.23. *(Optimization with column space constraint) Let A be a real* $m \times n$ *matrix. Let C be a* $n \times k$ *matrix. Fix* $b \in \mathbb{R}^m$. *We want to find a vector x that solves*

$$\begin{aligned} &\textit{Minimize:} \quad ||Ax - b||^2 \\ &\textit{Subject to:} \quad x \in Col(C) \end{aligned}$$

a) Find an optimal x^* *in terms of A, C, b. Hint: Convert to a least squares problem* $Hz \approx y$ *for some H, y. For simplicity, you can assume A and C both have linearly independent columns.*

b) Find a numerical value of $x^* \in \mathbb{R}^4$ *for the special case*

$$A = \begin{bmatrix} 7 & 0 & -3 & 1 \\ 0 & 4 & 3 & 3 \\ 8 & 1 & -5 & -7 \\ 0 & 0 & 9 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} -8 & 2 \\ 3 & 0 \\ 3 & 3 \\ 1 & 9 \end{bmatrix}, \quad b = \begin{bmatrix} 5 \\ 10 \\ 5 \\ 10 \end{bmatrix}$$

Make sure to verify that $x^* \in Col(C)$. *What is the optimal objective value* $||Ax^* - b||^2$?

EXERCISE 9.24. *(Optimization with linear equality constraints) Let A be a real* $m \times n$ *matrix. Let C be a* $k \times n$ *matrix with* $rank(C) = r$ *(assume* $r < n$*). Fix* $b \in \mathbb{R}^m$ *and* $d \in \mathbb{R}^k$. *We want to find*

an $x \in \mathbb{R}^n$ that solves the following optimization problem:

$$\begin{aligned} &\textit{Minimize:} && \|Ax - b\|^2 \\ &\textit{Subject to:} && Cx = d \end{aligned}$$

a) Let w be a particular *vector that satisfies $Cw = d$. Let M be a matrix with columns that consist of a basis for Null(C) (so M has $n - r$ columns). Show that $Cx = d$ if and only if $x = w + Mz$ for some $z \in \mathbb{R}^{n-r}$.*

b) Use w and M to find an x that solves the desired optimization problem. Hint: Reduce to a least squares problem $Hz \approx y$ for some matrix H and some vector y. If it helps, you can assume the matrix AM has linearly independent columns.

c) Solve the optimization problem to find the optimal x^* for the case when

$$A = \begin{bmatrix} 4 & -1 & 0 & 7 & 5 \\ 0 & 1 & 0 & 1 & -1 \\ 2 & 2 & 2 & 0 & 1 \\ 0 & 3 & -1 & -1 & 0 \end{bmatrix}, \quad C = \begin{bmatrix} 4 & 1 & 4 & 1 & 9 \\ 0 & 0 & 2 & -1 & 1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \quad d = \begin{bmatrix} 55 \\ 1 \end{bmatrix}$$

You can use $w = (1, 2, 0, 4, 5)$, but you will need to find a basis for Null(C). Make sure to verify that your x^* satisfies $Cx^* = d$. What is the optimal objective value $\|Ax^* - b\|^2$?

EXERCISE 9.25. *(Interpolation and overfit) Suppose we have 20 data samples $(1, y_1)$, $(2, y_2)$, ..., $(20, y_{20})$ for values $y_i \in \mathbb{R}$.*

a) Fix $v \in \mathbb{R}$ as an unusually large constant (such as $v = 10^{1000}$). Use Lemma 9.2 to show there is a degree-20 polynomial $p(t)$ that satisfies $p(i) = y_i$ for all $i \in \{1, ..., 20\}$, and $p(1.5) = v$. This interpolates the data but has unusually large function values in between the data points.

b) For these 20 data samples, suppose we use a polynomial model with degree less than or equal to 19, so $y_i \approx p(i)$ for $i \in \{1, ..., 20\}$ where $p(t) = \sum_{k=0}^{19} \theta_k t^k$. We use the least squares technique $y \approx A\theta$ to solve for the coefficients $\theta = (\theta_0, ..., \theta_{19})$ that minimize $\sum_{i=1}^{20} (y_i - p(i))^2$. Find the matrix A and the vector y. What is the size of A? Explain how we know the solution $\theta^* = (A^\top A)^{-1} A^\top y$ yields a polynomial $p^*(t) = \sum_{k=0}^{19} \theta_k^* t^k$ that satisfies $\sum_{i=1}^{20} (y_i - p^*(i))^2 = 0$. Explain why this zero-error result is not necessarily good.

EXERCISE 9.26. *(Nonlinear fit) We have data samples $(t_1, r_1, y_1), ..., (t_m, r_m, y_m)$, where $(t_i, r_i, y_i) \in \mathbb{R}^3$ for all $i \in \{1, ..., m\}$.*

a) Find coefficients θ_i that form a least squares fit to the following nonlinear model:

$$y_i \approx \theta_0 + \theta_1 \cos(t_i) + \theta_2 \cos(2t_i) + \theta_3 \cos(3t_i) + \theta_4 r_i \quad \forall i \in \{1, ..., m\}$$

b) How big is your A matrix? What condition on m makes A tall (more rows than columns)?

EXERCISE 9.27. *(Least squares fit for a nonlinear function of two variables) Consider a system that takes two variables (t, r) and ouputs a single variable y (with some noise):*

$$y = a + bt + cr + dtr + Noise$$

for unknown coefficients a, b, c, d. Assume we have m data points (t_i, r_i, y_i) for $i \in \{1, ..., m\}$. We want to do a standard least squares fit by choosing coefficients so that

$$y_i \approx a + bt_i + cr_i + dt_i r_i \quad \forall i \in \{1, ..., m\}$$

Describe the form of the A matrix and the y and θ vectors for this least squares $y \approx A\theta$ problem, assuming a general collection of data (t_i, r_i, y_i) for $i \in \{1, ..., m\}$. What is the size of A? What size of m makes A a tall matrix? Assuming A has linearly independent columns, what is the optimal (a^*, b^*, c^*, d^*)? What is the first equation of the system of equations $A^\top A\theta^* = A^\top y$?

EXERCISE 9.28. *(Least squares linear fit) We have m data samples $\{(r_i, t_i, y_i)\}_{i=1}^m$. We fit to a linear (affine) model*

$$y_i \approx \theta_0 + \theta_1 r_i + \theta_2 t_i \quad \forall i \in \{1, ..., m\}$$

Find $\theta^* = (\theta_0^*, \theta_1^*, \theta_2^*) \in \mathbb{R}^3$ to minimize $\sum_{i=1}^m (y_i - f(r_i, t_i))^2$ where $f(r, t) = \theta_0 + \theta_1 r + \theta_2 t$. What is the corresponding y vector and A matrix for the equation $y \approx A\theta$? How many samples m are needed to ensure matrix A is tall?

EXERCISE 9.29. *(Least squares with sines and cosines) We have m data samples $\{(r_i, t_i, y_i)\}_{i=1}^m$. We fit to a model*

$$y_i \approx \theta_0 + \theta_1 \sin(r_i) + \theta_2 \cos(t_i) \quad \forall i \in \{1, ..., m\}$$

Find $\theta^* = (\theta_0^*, \theta_1^*, \theta_2^*) \in \mathbb{R}^3$ to minimize $\sum_{i=1}^m (y_i - f(r_i, t_i))^2$ where $f(r, t) = \theta_0 + \theta_1 \sin(r) + \theta_2 \cos(t)$. What is the corresponding y vector and A matrix for the equation $y \approx A\theta$? How many samples m are needed to ensure matrix A is tall?

EXERCISE 9.30. *(Another least squares) We have m data samples $\{(r_i, t_i, y_i)\}_{i=1}^m$. We fit to a model*

$$y_i \approx \theta_1 + \theta_2 t_i e^{r_i} \quad \forall i \in \{1, ..., m\}$$

Find $(\theta_1^*, \theta_2^*) \in \mathbb{R}^2$ to minimize $\sum_{i=1}^m (y_i - f(r_i, t_i))^2$ where $f(r, t) = \theta_1 + \theta_2 t e^r$. What is the corresponding y vector and A matrix for the equation $y \approx A\theta$? How many samples m are needed to ensure matrix A is tall?

EXERCISE 9.31. *(Projections with pesky dependencies) Consider*

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 0 & -1 & -1 \\ 2 & 2 & 2 \\ 4 & 0 & 0 \end{bmatrix}, \quad y = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

a) Project y onto $Col(A)$. Then complete the paragraph: In general, suppose we want to project a vector $y \in \mathbb{R}^m$ onto the column space of a real $m \times n$ matrix A. If A does not have linearly independent columns then we can always...

b) Complete the paragraph: In a general real inner product space V, suppose we want to project a vector $y \in V$ onto $Span\{v_1, ..., v_k\}$. If the Gram matrix G is not invertible then we can always...

EXERCISE 9.32. *(Least squares linear fit) Suppose we have m data points (t_i, y_i) for $i \in \{1, ..., m\}$. We want to find optimal coefficients a, b to fit the data to the following linear model (this is often called* linear regression*):*

$$y_i \approx at_i + b \quad \forall i \in \{1, ..., m\}$$

a) Write this as a $y \approx Ax$ problem for some column vectors $y = (y_1; \ldots; y_m)$ and $x = (a; b)$, and some $m \times 2$ matrix A.

b) Show that

$$A^\top A = \begin{bmatrix} \sum_{i=1}^m t_i^2 & \sum_{i=1}^m t_i \\ \sum_{i=1}^m t_i & m \end{bmatrix}, \quad A^\top y = \begin{bmatrix} \sum_{i=1}^m t_i y_i \\ \sum_{i=1}^m y_i \end{bmatrix}$$

c) Define empirical averages $\overline{y}, \overline{t}, \overline{ty}, \overline{t^2}$ by

$$\overline{y} = \frac{1}{m}\sum_{i=1}^m y_i, \quad \overline{t} = \frac{1}{m}\sum_{i=1}^m t_i, \quad \overline{ty} = \frac{1}{m}\sum_{i=1}^m t_i y_i, \quad \overline{t^2} = \frac{1}{m}\sum_{i=1}^m t_i^2$$

Show that the least squares matrix equation $A^\top Ax = A^\top y$ is equivalent to two linear equations with two unknowns a and b and with coefficients $\overline{y}, \overline{t}, \overline{ty}, \overline{t^2}$. From your equations, verify that the optimal a and b must satisfy $\overline{y} = a\overline{t} + b$.

EXERCISE 9.33. *(Least squares with vector output) We have m data samples $\{(r_i, t_i, y_i, z_i)\}_{i=1}^m$. We fit to a model*

$$\begin{bmatrix} y_i \\ z_i \end{bmatrix} \approx \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} r_i \\ t_i \end{bmatrix} \quad \forall i \in \{1, ..., m\}$$

Using the standard Euclidean norm for $\mathbb{R}^2$, find $(a, b, c, d) \in \mathbb{R}^4$ to minimize

$$\sum_{i=1}^{m} \left\| \begin{bmatrix} y_i \\ z_i \end{bmatrix} - \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} r_i \\ t_i \end{bmatrix} \right\|^2$$

EXERCISE 9.34. *(Least squares with vertical translation) Suppose we have data $(\vec{z}_i, y_i)$ for $i \in \{1, \ldots, m\}$, where $\vec{z}_i \in \mathbb{R}^n$ and $y_i \in \mathbb{R}$ for all i. We have a candidate function $f : \mathbb{R}^n \to \mathbb{R}$ such that*

$$y_i \approx f(\vec{z}_i) \quad \forall i \in \{1, \ldots, m\}$$

which has error

$$error = \sum_{i=1}^{m} (y_i - f(\vec{z}_i))^2$$

Suppose that $\frac{1}{m}\sum_{i=1}^{m} f(\vec{z}_i) \neq \frac{1}{m}\sum_{i=1}^{m} y_i$, so the empirical average of the y_i data is not *the same as the empirical average of $f(\vec{z}_i)$. Show that we can strictly reduce error by using a new function $\tilde{f} : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$\tilde{f}(z) = f(z) + a_0 \quad \forall z \in \mathbb{R}^n$$

for some $a_0 \in \mathbb{R}$ such that the new function satisfies $\frac{1}{m}\sum_{i=1}^{m} \tilde{f}(\vec{z}_i) = \frac{1}{m}\sum_{i=1}^{m} y_i$. How much is the error reduced?

EXERCISE 9.35. *(Computer fit to quadratic model) Consider the following input-output system model:*

$$y = a + bt + ct^2 + N \quad , \quad t \in [-1.8, 1.8]$$

where t is the input, N is zero-mean Gaussian noise, and y is the output. The true coefficients a, b, c are unknown. The variance σ^2 of the noise is also unknown. We want to learn a, b, c from m data points (t_i, y_i) for $i \in \{1, \ldots, m\}$, so that

$$y_i = a + bt_i + ct_i^2 + N_i \quad \forall i \in \{1, \ldots, m\}$$

where $\{N_1, \ldots, N_m\}$ are i.i.d. Gaussian with mean zero and variance σ^2.

a) Explain how to compute best-fit coefficients a_m^*, b_m^*, c_m^* based on the m data points (t_i, y_i) for $i \in \{1, \ldots, m\}$

b) Write a computer program to generate m data points under the assumptions:

$$a_{true} = 1.37$$
$$b_{true} = -2.92$$
$$c_{true} = 4.63$$
$$\sigma^2 = 25.0$$

You should choose t_i independently and uniformly over $[-1.8, 1.8]$ for $i \in \{1, \ldots, m\}$. This obtains the (t_i, y_i) data. Now, find the best-fit approximations a_m^*, b_m^*, c_m^* for: (i) $m = 10$ data points; (ii) $m = 20$ data points; (iii) $m = 200$ data points. Are the approximations getting closer to $(a_{true}, b_{true}, c_{true})$?

c) Plot your curves to show the true model; the data; the Fit-10 approximation; the Fit-20 approximation; the Fit-200 approximation.

d) Try using more data points $m > 200$ to get more accuracy between (a_m^*, b_m^*, c_m^*) and $(a_{true}, b_{true}, c_{true})$.

EXERCISE 9.36. *(Weighted least squares) Define $y = (1, 2, 3, 4)$.*

a) Find $p = (p_1, p_2, p_3, p_4) \in Span\{(1, 1, 1, 1)\}$ to minimize $\sum_{i=1}^4 (y_i - p_i)^2$. Does the orthogonality condition hold?

b) Find $q = (q_1, q_2, q_3, q_4) \in Span\{(1, 1, 1, 1)\}$ to minimize $16(y_1 - q_1)^2 + \sum_{i=2}^4 (y_i - q_i)^2$. Compare to part (a). Show that the orthogonality condition no longer holds if the standard inner product is used, but the orthogonality condition holds if a new inner product is used.

EXERCISE 9.37. *(Weighted least squares design) Design and solve your own weighted least squares problem. You should compare weighted and unweighted solutions.*

EXERCISE 9.38. *(Finding a matrix) Using the standard inner product for $\mathbb{R}^3$, find a matrix A such that the projection of $(1, 3, 3)$ onto $Col(A)$ is $(-1, 1, 3)$.*

EXERCISE 9.39. *(Information from projections) We have a real matrix A. Using standard inner products in $\mathbb{R}^3$, we find that projecting $(33, 15, 0)$ onto $Col(A)$ gives $(1, -1, 4)$. Projecting $(56, 30, 1)$ onto $Col(A)$ gives $(0, 2, 8)$.*

a) What is the projection of $(89, 45, 1)$ onto $Col(A)$?

b) What is the projection of $(53, 15, -2)$ onto $Col(A)$?

c) What is the projection of $(-21, -37, 7)$ onto $Col(A)$?

d) Show that $rank(A) = 2$.

e) Find a matrix A with these properties.

EXERCISE 9.40. *(Fourier series: Three cosines and a constant) Let V be the real vector space of square integrable functions $f : [-1/2, 1/2] \to \mathbb{R}$. Use the standard inner product $\langle f, g \rangle = \int_{-1/2}^{1/2} f(t)g(t)dt$ for $f, g \in V$. Define function $y(t) \in V$ by*

$$y(t) = \min\{1, 4|t|\} \quad \forall t \in [-1/2, 1/2]$$

The function $y(t)$ is illustrated in Fig. 9.13. It has three sharp points where it is not differentiable. We want to project $y(t)$ onto $Span\{1, \sqrt{2}\cos(2\pi t), \sqrt{2}\cos(4\pi t), \sqrt{2}\cos(6\pi t)\}$. We know the projection function $p \in V$ has the form

$$p(t) = a + b\sqrt{2}\cos(2\pi t) + c\sqrt{2}\cos(4\pi t) + d\sqrt{2}\cos(6\pi t) \quad \forall t \in [-1/2, 1/2]$$

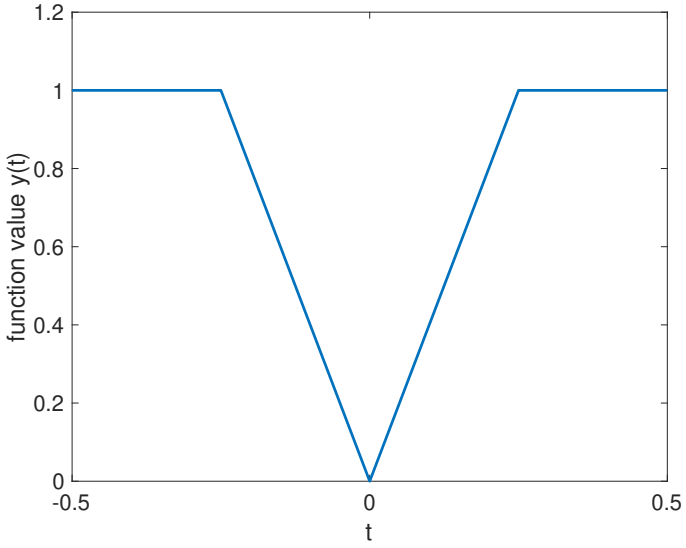


FIGURE 9.13. The function $y(t) = \min\{1, 4|t|\}$ over the interval $[-1/2, 1/2]$ for Exercise 9.40.

for some real numbers a, b, c, d. Of course, $p(t)$ is differentiable everywhere. This problem shows what happens when we approximate a nonsmooth function using a linear combination of smooth functions.

a) Find a, b, c, d.

b) Using a computer (such as MATLAB*), plot $y(t)$ and the projection $p(t)$ over $t \in [-1/2, 1/2]$.*

c) Prove that $y(t)$ is an even function*, meaning that $y(-t) = y(t)$ for all $t \in [-1/2, 1/2]$.*

EXERCISE 9.41. *(Two cosines, two sines, and a constant) Let V be the same real inner product space of Exercise 9.40. Define $y(t) \in V$ by $y(t) = \max\{0, t\}$ for $t \in [-1/2, 1/2]$.*

a) Find the projection of $y(t)$ onto $Span\{1, \sqrt{2}\cos(2\pi t), \sqrt{2}\cos(4\pi t), \sqrt{2}\sin(2\pi t), \sqrt{2}\sin(4\pi t)\}$.

b) Plot $y(t)$ and its projection over the interval $t \in [-1/2, 1/2]$.

EXERCISE 9.42. *(Odd and even functions) A function $f : \mathbb{R} \to \mathbb{R}$ is defined to be* even *if $f(-t) = f(t) \quad \forall t \in \mathbb{R}$. The function is* odd *if $f(-t) = -f(t) \quad \forall t \in \mathbb{R}$. For example, given a constant $\theta \neq 0$, the function $\cos(\theta t)$ is even, while $\sin(\theta t)$ is odd.*

a) Show that every function $f : \mathbb{R} \to \mathbb{R}$ can be written as the sum of an even function and an odd function by

$$f(t) = \left(\frac{f(t)+f(-t)}{2}\right) + \left(\frac{f(t)-f(-t)}{2}\right) \quad \forall t \in \mathbb{R}$$

Specifically, show the first component is even (called the even part *of the function), and the other is odd (called the* odd part*).*

b) Show the decomposition in part (a) is unique: Suppose f is a function that satisfies $f(t) = h(t) + g(t)$ for all $t \in \mathbb{R}$, where h is even and g is odd. Show that we must have $h(t) = \frac{1}{2}(f(t) + f(-t))$ and $g(t) = \frac{1}{2}(f(t) - f(-t))$.

EXERCISE 9.43. *(Integral of odd and even functions) Show that if f, g are square integrable functions and f is even, g is odd, then for any $a > 0$ we have $\int_{-a}^{a} f(t)g(t)dt = 0$. Hint: use $\int_{-a}^{a} = \int_{-a}^{0} + \int_{0}^{a}$ and for one of the integrals use a change of variable $u = -t$.*

EXERCISE 9.44. *(Fourier series) Let V be the real vector space of square integrable functions $f : [-1/2, 1/2] \to \mathbb{R}$ with inner product $\langle f, g\rangle = \int_{-1/2}^{1/2} f(t)g(t)dt$ for all $f, g \in V$. Let n be a positive integer. Define the subspaces $S_n \subseteq V$ and $G_n \subseteq V$ by*

$$S_n = Span\{\sqrt{2}\sin(2\pi kt)\}_{k=1}^{n}$$
$$G_n = Span\{1, \{\sqrt{2}\cos(2\pi kt)\}_{k=1}^{n}, \{\sqrt{2}\sin(2\pi kt)\}_{k=1}^{n}\}$$

a) Let $h \in V$ be an even function. Find the projection of h onto the subspace S_n.

b) Let $h \in V$ be an even function. Let $p_n \in V$ denote the projection of h onto the subspace G_n. Show

$$p_n(t) = a_0 + \sum_{k=1}^{n} a_k\sqrt{2}\cos(2\pi kt) + \sum_{k=1}^{n} b_k\sqrt{2}\sin(2\pi kt)$$

where $a_0 = \langle h, 1\rangle$ and $a_k = \langle h, \sqrt{2}\cos(2\pi kt)\rangle$, $b_k = \langle h, \sqrt{2}\sin(2\pi kt)\rangle$ for $k \in \{1, \ldots, n\}$. Compute the numerical b_k values.

EXERCISE 9.45. *(Energy) Consider the inner product space V of the previous problem. Fix $h(t) \in V$ and let $p_n(t)$ be its projection onto the subspace G_n. The definition of p_n and G_n are in Exercise 9.44. Define the error function $e_n(t) = h(t) - p_n(t)$. Define $||h||^2 = \int_{-1/2}^{1/2} h(t)^2 dt$ and note that, since $h \in V$, we know $||h||^2 < \infty$.*

a) Show that $||h||^2 = ||p_n||^2 + ||e_n||^2$.

b) Give an expression for $||e_n||^2$ in terms of $||h||^2$ and the coefficients $a_0, a_1, \ldots, a_n, b_1, \ldots, b_n$.

EXERCISE 9.46. *(Dumbledore at Hogwarts) Professor Dumbledore was playing around with magical theory and decided to mix a standard summoning spell with a remembrall. There was a mistake in his calculations and, quite accidentally, he ended up magically adding two more chimneys onto the rooftop of Hogwarts castle (compare the old Hogwarts in Fig. 9.8 with the new one in Fig. 9.14).*[4] *While Hogwarts formerly had only 2 chimneys, the new 4-chimney Hogwarts*

[4] Dumbledore's mistake was noticed in a parallel universe by Mr. Norrell, who declared it to be a poor example of English Magic that would never have happened if Dumbledore were a more practical magician. Meanwhile, Jonathan Strange excitedly observed that the new Hogwarts was an exact replica of the long-lost home of the Raven King.

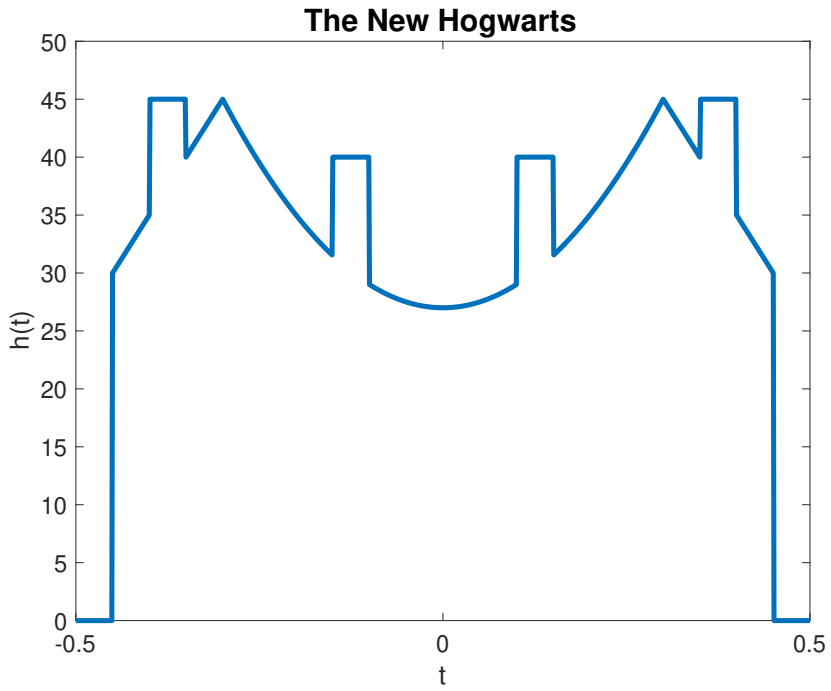


FIGURE 9.14. The new 4-chimney Hogwarts castle for Exercise 9.46.

is the following even function $h : [-1/2, 1/2] \to \mathbb{R}$ *defined (using* $b = 30$*,* $c = 75$*) by*

$$h(t) = \begin{cases} 0 & \text{if } |t| > 0.45 \\ 45 & \text{if } |t| \in (0.35, 0.4) \\ c - 100|t| & \text{if } |t| \in (.3, .35] \cup [.4, .45] \\ 40 & \text{if } |t| \in (.1, .15) \\ -(2/100)b^2 + c - b + 200t^2 & \text{else} \end{cases}$$

a) Using numerical integration, find $\|h\|^2 = \int_{-1/2}^{1/2} h(t)^2 dt$*, the energy of the new Hogwarts.*

b) Make plots of: (i) The $h(t)$ *function with* $p_3(t)$*; (ii) The* $h(t)$ *function with* $p_{10}(t)$*; (iii) The* $h(t)$ *function with* $p_{50}(t)$*; (iv) The* $h(t)$ *function with* $p_{200}(t)$*. Here,* $p_n(t)$ *is defined as in Exercise 9.44.*

c) Find $\|e_3\|^2$*,* $\|e_{10}\|^2$*,* $\|e_{50}\|^2$*,* $\|e_{200}\|^2$ *using the energy result of the previous problem. You will need to use decimal precision of at least four or five decimal places. Here,* e_n *is defined as in Exercise 9.45.*

EXERCISE 9.47. *(Gram matrix for random variables) Suppose we independently flip a fair coin three times. Define random variable* X *to be the number of heads. Define random variables* Y, Z

by

$$Y = \begin{cases} 1 & \text{if flip 1 is heads} \\ 0 & \text{else} \end{cases}, \quad Z = \begin{cases} 1 & \text{if flip 2 is heads} \\ 0 & \text{else} \end{cases}$$

Thus, $X \in \{0,1,2,3\}$, $Y, Z \in \{0,1\}$. Let V be the real inner product space of random variables with finite second moments. We use the inner product $\langle R, S\rangle = \mathbb{E}[RS]$ for all $R, S \in V$. Note that $X, Y, Z \in V$.

a) Find the 3×3 Gram matrix G for $\{X, Y, Z\}$. Hint: Observe $X = Y + Z + W$ for some Bernoulli random variable W.

b) Verify that $\{X, Y, Z\}$ are linearly independent by showing G is invertible. What is G^{-1}?

EXERCISE 9.48. *(Projections for random variables) Consider the random variables X, Y, Z in Exercise 9.47. Using the standard inner product $\langle R, S\rangle = \mathbb{E}[RS]$ for random variables R, S with finite second moments, do:*

a) Project X onto $Span\{1, Y, Z\}$. Is the resulting projection $\hat{X}$ unbiased, so that $\mathbb{E}[\hat{X}] = \mathbb{E}[X]$?

b) Project X onto $Span\{Y, Z\}$. Is the resulting projection $\hat{X}$ unbiased? Compare with part (a).

c) Project Y onto $Span\{X, Z\}$. Then, project Z onto $Span\{X, Y\}$.

EXERCISE 9.49. *(Gram-Schmidt for random variables) Let X, Y, Z be the random variables of Exercise 9.47. Treat these as vectors in the vector space V of random variables with finite second moments. Use inner product $\langle R, S\rangle = \mathbb{E}[RS]$ for all $R, S \in V$. Use the Gram-Schmidt procedure on $\{X, Y, Z\}$ to get orthonormal random variables $\{R_1, R_2, R_3\}$ with the same span as $\{X, Y, Z\}$.*

EXERCISE 9.50. *(Decorrelating) Consider a probability space $(\Omega, \mathcal{F}, P)$. Let V be the vector space of random variables on this probability space that have zero mean and finite second moments. Use inner product $\langle X, Y\rangle = \mathbb{E}[XY]$ for all $X, Y \in V$. Let X, Y, Z be three random variables in V and let G be the corresponding 3×3 Gram matrix. We know G is positive semidefinite, and is positive definite if and only if $\{X, Y, Z\}$ are linearly independent when treated as vectors in V. Assume G is positive definite (hence invertible) and let $G^{1/2}$ denote a 3×3 matrix that satisfies $G^{1/2}G^{1/2} = G$. Such a matrix $G^{1/2}$ can be viewed as a "square root" matrix.*[5]

a) Argue that $G^{1/2}$ is invertible and let $G^{-1/2}$ denote its inverse. Argue that $G^{-1/2}G^{-1/2} = G^{-1}$.

b) Define random variables Q, R, S by

$$\begin{bmatrix} Q \\ R \\ S \end{bmatrix} = G^{-1/2} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix}$$

[5] It can be shown that such a "square root" exists for any symmetric and positive semidefinite matrix G. It can be computed via a *singular value decomposition* of G. In particular, $G = U\Lambda U^T$ and $G^{1/2} = U\Lambda^{1/2}U^T$, where $U^T U = I$ and Λ is purely diagonal with nonnegative diagonal entries. The square root is invertible if and only if G is invertible. Also, G is invertible if and only if all diagonal entries of Λ are positive.

Show that Q, R, S are pairwise uncorrelated, and each has second moment 1, by computing the Gram matrix of $\{Q, R, S\}$, which is the expectation of the following random 3×3 matrix:[6]

$$\begin{bmatrix} Q \\ R \\ S \end{bmatrix} \begin{bmatrix} Q & R & S \end{bmatrix}$$

EXERCISE 9.51. *(MMSE) Use* (9.37) *to show that if $\mathbb{E}[X] = 0$ and $\mathbb{E}[Y_i] = 0$ for all i, then $\theta_0 = 0$. In this case, if $n = 3$, show that the solution is $\hat{X} = \sum_{i=1}^{3} \theta_i Y_i$ where $(\theta_1, \theta_2, \theta_3)$ solves*

$$\begin{bmatrix} Cov(X, Y_1) \\ Cov(X, Y_2) \\ Cov(X, Y_3) \end{bmatrix} = \begin{bmatrix} Var(Y_1) & Cov(Y_1, Y_2) & Cov(Y_1, Y_3) \\ Cov(Y_1, Y_2) & Var(Y_2) & Cov(Y_2, Y_3) \\ Cov(Y_1, Y_3) & Cov(Y_2, Y_3) & Var(Y_3) \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \\ \theta_3 \end{bmatrix}$$

Argue that the 3×3 matrix in the above equality is the Gram matrix for $\{Y_1, Y_2, Y_3\}$ and is not invertible if and only if either $Y_1 \equiv 0$, Y_2 is equivalent to a vector in $Span(Y_1)$, or Y_3 is equivalent to a vector in $Span\{Y_1, Y_2\}$.

EXERCISE 9.52. *(Another MMSE) Find the optimal θ_i coefficients for the linear estimator $\hat{X} = \theta_0 + \sum_{i=1}^{n} \theta_i Y_i$ when observations Y_i are given by the model of* (9.38)*. Here, $\rho_1, \ldots, \rho_n$ are known real numbers. Also, for each $i \in \{1, \ldots, n\}$, N_i is independent noise that is known to have zero mean and nonzero variance σ_i^2.*

EXERCISE 9.53. *(Projections shrink distances) Let V be a real inner product space. Suppose S is a finite dimensional subspace of V.*

a) Fix $y \in V$. Let $p = Proj_S(y)$. Show that $\|y\| \geq \|p\|$. Hint: Use the orthogonality conditions with $y = p + (y - p)$.

b) Fix $y_1, y_2 \in V$. Define $p_i = Proj_S(y_i)$ for $i \in \{1, 2\}$. Show that $\|y_1 - y_2\| \geq \|p_1 - p_2\|$.

c) The results of (a) and (b) only assume $\dim(S) < \infty$ to ensure existence *of the projections. Explain why the results still hold when $\dim(S) = \infty$ if we assume the projections exist.*

EXERCISE 9.54. *(Cosines with noninteger frequencies) Let V be the inner product space of Exercise 9.40. As in that problem, let $y(t) = \min\{1, 4|t|\}$ for $t \in [-1/2, 1/2]$.*

a) Project $y(t)$ onto

$$Span\{1, \sqrt{2}\cos(2\pi t), \sqrt{2}\cos(4\pi t), \sqrt{2}\cos(6\pi t), \sqrt{2}\cos(8\pi t), \cos(2.5\pi t)\}$$

Note that these functions are no longer orthonormal, so you will need to use the 6×6 Gram matrix G. Make sure to specify the Gram matrix G. Also, plot the results. Compare to the projection

[6] From a probability theory perspective, it can be shown that if (X, Y, Z) are *jointly Gaussian* then the decorrelated random variables (Q, R, S) are jointly Gaussian and mutually independent.

without the $\cos(2.5\pi t)$ *and make a conjecture about what is happening.*

b) Replace the $\cos(2.5\pi t)$ *function with* $\cos(10.5\pi t)$ *and repeat. Make sure to specify the new* 6×6 *Gram matrix.*

EXERCISE 9.55. *(Projecting onto Null*(A)*) Consider*

$$A = \begin{bmatrix} 2 & 3 & -1 & 0 \\ 0 & 1 & 0 & 2 \\ 3 & 0 & 2 & 1 \end{bmatrix} \quad , \quad w = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$$

a) Show that $w \notin Null(A)$.

b) Define p *as the projection of* w *onto Null*(A). *Find* p *using* (9.46). *Verify* $p \in Null(A)$.

c) Use Gaussian elimination to find another point in Null(A) *and compare its distance to* w *with* $||w - p||$.

d) What is the projection of $(42, 1, 3)$ *onto Null*$(A^\top)$*?*

EXERCISE 9.56. *(Another projecting onto Null*(A)*) Design your own problem related to projecting onto Null*(A).

EXERCISE 9.57. *(Projecting on an affine set) We want a vector* $x \in \mathbb{R}^5$ *to solve*

$$\begin{aligned} &\textit{Minimize:} && \textstyle\sum_{i=1}^{5}(x_i - i)^2 \\ &\textit{Subject to:} && Ax = b \end{aligned}$$

$$A = \begin{bmatrix} 3 & -1 & 3 & 4 & 0 \\ 2 & 2 & 2 & 0 & 2 \\ 1 & -1 & 1 & -1 & 1 \end{bmatrix} \quad , \quad b = \begin{bmatrix} -10 \\ 22 \\ -3 \end{bmatrix}$$

a) Find any particular vector $x_p \in \mathbb{R}^5$ *that solves* $Ax = b$. *You can use either Gaussian elimination or Lemma 7.1.*

b) Show how the optimization problem of interest is equivalent to projecting a particular point $c \in \mathbb{R}^5$ *onto Null*(A). *In terms of your* x_p *vector, what is the vector* c*?*

c) Using part (b) with the formula (9.46)*, find a vector* x^* *that solves the optimization problem of interest. Verify it is the same as the answer obtained by formula* (9.51) *using* $w = (1, 2, 3, 4, 5)$.

EXERCISE 9.58. *(Properties of projection function) Let* V *be a real inner product space. Let* S *be a finite dimensional subspace of* V. *Suppose each* $y \in V$ *has a closest point in* S *(Theorem 9.1 ensures this closest point is unique). Let* $P : V \to V$ *be the linear operator that maps each point* $y \in V$ *to its unique closest point in* S. *We want to show* $\langle P(x), y\rangle = \langle x, P(y)\rangle$ *for all* $x, y \in V$.

a) Argue that $\langle x - P(x), s\rangle = 0$ *for all* $x \in V, s \in S$.

b) Use part (a) to show $\langle x, P(y)\rangle = \langle P(x), P(y)\rangle$ for all $x, y \in V$.

c) Explain why $\langle y, P(x)\rangle = \langle P(y), P(x)\rangle$ for all $x, y \in V$. Use this to prove the desired result.

EXERCISE 9.59. *(Projection operator) This exercise relates to Lemma 9.6. Suppose V is a real inner product space and $\dim(V)$ is a positive integer. Let $P : V \to V$ be a projection operator. Define $S = Span(P)$. Let $\{v_1, ..., v_k\}$ be a basis for S. Let $\{y_1, ..., y_m\}$ be additional vectors such that $\{v_1, ..., v_k, y_1, ..., y_m\}$ is a basis for V. Show $\{v_1, ..., v_k, y_1 - P(y_1), ..., y_m - P(y_m)\}$ is linearly independent.*

EXERCISE 9.60. *(Property of projection operator) Let V be a vector space over a field $\mathbb{F}$. Suppose $P : V \to V$ is a projection operator (defined in Section 9.16.2). Define $S = Span(P)$. Show that $s \in S$ if and only if $P(s) = s$.*

EXERCISE 9.61. *(Projecting onto an infinite dimensional space) Let V be the real vector space of square summable sequences $(x_1, x_2, x_3, ...) \in \mathbb{R}^{\mathbb{N}}$. Use the standard inner product $\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i$ for $x, y \in V$. Define $S = \{(x_1, x_2, ...) \in V : x_1 = 0\}$.*

a) Prove S is a subspace. Then, prove $\dim(S) = \infty$.

b) For each $y = (y_1, y_2, y_3, ...) \in V$, find the projection of y onto S. Explain why.

EXERCISE 9.62. *(Projecting onto the even functions) Let V be the real vector space of functions $f : [-1, 1] \to \mathbb{R}$ that are continuous (such functions are square integrable and have the property $(f \equiv g) \iff (f = g)$, see final paragraph of Chapter 8.6). Use the standard inner product $\langle f, g\rangle = \int_{-1}^{1} f(t)g(t)dt$. Define the subsets of* even *and* odd *functions:*

$$S_{even} = \{f \in V : f(-t) = f(t) \quad \forall t \in [-1, 1]\}$$
$$S_{odd} = \{f \in V : f(-t) = -f(t) \quad \forall t \in [-1, 1]\}$$

From Exercise 9.43 we know that if $f \in S_{even}$ and $g \in S_{odd}$ then $\langle f, g\rangle = 0$.

a) Show that S_{even} and S_{odd} are subspaces.

b) Show that $\dim(S_{even}) = \infty$ and $\dim(S_{odd}) = \infty$.

c) Define $T_{even} : V \to V$ and $T_{odd} : V \to V$ by mapping each $f \in V$ as follows:

$$T_{even}(f)(t) = \frac{f(t) + f(-t)}{2} \quad \forall t \in [-1, 1]$$
$$T_{odd}(f)(t) = \frac{f(t) - f(-t)}{2} \quad \forall t \in [-1, 1]$$

Show that for each $f \in V$, $T_{even}(f)$ is the projection of f onto S_{even}.

EXERCISE 9.63. *Suppose $\{X, N_1, ..., N_n\}$ are random variables with zero mean and finite variances. Suppose they are pairwise uncorrelated (so $\mathbb{E}[XN_i] = 0$ for all i, and $\mathbb{E}\left[N_i N_j\right] = 0$ for $i \neq j$). Thus, when treated as vectors in the real inner product space V of random variables with*

finite second moments, $\{X, N_1, \ldots, N_n\}$ *are pairwise orthogonal. Define*

$$Y_i = \rho_i X + N_i \quad \forall i \in \{1, \ldots, n\}$$

Let $\sigma_i^2 = \mathbb{E}\left[N_i^2\right]$ *for* $i \in \{1, \ldots, n\}$*. Suppose* $\sigma_i^2 > 0$ *for all* $i \in \{1, \ldots, n\}$*.*

a) Show Y_1 *is not equivalent to* 0*, meaning that* $\mathbb{E}\left[Y_1^2\right] > 0$*.*

b) Fix $i \in \{2, \ldots, n\}$*. Let* $c_1, \ldots, c_{i-1}$ *be any real-valued constants. Show that* Y_i *is not equivalent to* $\sum_{j=1}^{i-1} c_j Y_j$*. Specifically, show* $\mathbb{E}\left[\left(Y_i - \sum_{j=1}^{i-1} c_j Y_j\right)^2\right] > 0$*.*

c) Conclude that $\{Y_1, \ldots, Y_n\}$ *is linearly independent (as vectors in the vector space* V*) and so the* $n \times n$ *Gram matrix* $G = \left(\mathbb{E}\left[Y_i Y_j\right]\right)$ *is invertible. [Note here that "linearly independent in the vector space" is not the same as "mutually independent in the probability space."]*

CHAPTER 10.

Complex inner product space

This is a brief chapter that treats Hermitians and complex inner product spaces. Throughout this chapter we use the field of complex numbers. So $\mathbb{F} = \mathbb{C}$. Complex numbers are useful for the theory of eigenvalues (Chapter 12). They are also used for Fourier series and Fourier transforms. They are extensively used in applications of signal processing, data communications, and quantum physics.

10.1. Complex conjugate

Every complex number $z \in \mathbb{C}$ has the form $z = a + bi$ for $a, b \in \mathbb{R}$. The real number a is called the *real part of* z; the real number b is called the *imaginary part of* z. We define the *complex conjugate of* z, written $\overline{z}$, as $a - bi$:

$$z = a + bi$$
$$\overline{z} = a - bi$$

For simplicity, the complex conjugate of z is often called the conjugate of z. Observe that $z = \overline{z}$ if and only if $z \in \mathbb{R}$. Also

$$z + \overline{z} = 2a \quad \text{(purely real)}$$
$$z - \overline{z} = 2bi \quad \text{(purely imaginary)}$$

Note that $\overline{z}z$ is purely real, and is nonnegative, because

$$\begin{aligned}\overline{z}z &= (a - bi)(a + bi)\\ &= a^2 - b^2 i^2\\ &= a^2 + b^2\end{aligned}$$

where the last equality uses $i^2 = -1$. We define the *magnitude* of z, written $|z|$, as the square root of $\overline{z}z$:

$$|z| = \sqrt{\overline{z}z}$$

For each $z \in \mathbb{C}$, the magnitude $|z|$ is real, nonnegative, and is zero if and only if $z = 0$.

Two important properties of complex conjugates are: (i) The conjugate of a sum of complex numbers is the sum of conjugates; (ii) The conjugate of a product is a product of conjugates. Indeed, if $z_1, z_2, \ldots, z_k \in \mathbb{C}$ then

$$\begin{aligned}\overline{z_1 + z_2 + \ldots + z_k} &= \overline{z}_1 + \overline{z}_2 + \ldots + \overline{z}_k\\ \overline{z_1 z_2 \cdots z_k} &= \overline{z}_1 \overline{z}_2 \cdots \overline{z}_k\end{aligned}$$

It follows that if $z_1, \ldots, z_k$ and $w_1, \ldots, w_k$ are complex numbers then

$$\overline{\sum_{j=1}^{k} z_j w_j} = \sum_{j=1}^{k} \overline{z}_j \overline{w}_j \tag{10.1}$$

10.2. Hermitian: The conjugate transpose

Let A be a $m \times n$ matrix with entries in $\mathbb{C}$. Define $\overline{A}$ as the matrix formed by taking complex conjugates of every entry of A. For example:

$$A = \begin{bmatrix} (3+4i) & i & 0 \\ -1 & (2-i) & 5i \end{bmatrix} \Longrightarrow \overline{A} = \begin{bmatrix} (3-4i) & -i & 0 \\ -1 & (2+i) & -5i \end{bmatrix}$$

Using (10.1), it can be shown that the complex conjugate of a product of matrices is the product of their complex conjugates:

$$\overline{A_1 A_2 \cdots A_k} = \overline{A}_1 \overline{A}_2 \cdots \overline{A}_k \tag{10.2}$$

Let A be a $m \times n$ matrix with entries in $\mathbb{C}$. Define A^H, read "A Hermitian," as the conjugate transpose of A. It does not matter if we take the transpose first, then take

complex conjugates, or the other way round:

$$A^H = \overline{A}^\top = \overline{A^\top}$$

For example

$$A = \begin{bmatrix} (3+4i) & i & 0 \\ -1 & 2-i & 5i \end{bmatrix} \implies A^H = \begin{bmatrix} (3-4i) & -1 \\ -i & 2+i \\ 0 & -5i \end{bmatrix}$$

From (10.2), it is an easy exercise to show that the Hermitian of a multiplication of matrices is the multiplication of their Hermitians in reverse order:

$$(A_1 A_2 \cdots A_{k-1} A_k)^H = A_k^H A_{k-1}^H \cdots A_2^H A_1^H$$

A matrix A is said to be *Hermitian* if it is equal to its conjugate transpose:

$$A \text{ is Hermitian} \iff A^H = A$$

In other words, a matrix is Hermitian if it has *conjugate symmetry*. Clearly every Hermitian matrix is square. In the world of complex numbers, conjugate symmetry is often more important than basic symmetry. Theoretical results that hold for real-valued symmetric matrices typically can be extended to complex-valued Hermitian matrices.

10.3. Complex inner product

Let V be a vector space over the field $\mathbb{C}$. A function that maps two vectors $x, y \in V$ to a complex number $\langle x, y \rangle \in \mathbb{C}$ is said to be a *complex inner product* if the following three axioms hold:[1]

1. Nonnegativity: For all $x \in V$ we have $\langle x, x \rangle \in \mathbb{R}$ and

   $$\langle x, x \rangle \geq 0$$

   Further, $\langle x, x \rangle = 0$ if and only if $x = 0$.

2. Conjugate symmetry: For all $x, y \in V$ we have

   $$\langle y, x \rangle = \overline{\langle x, y \rangle}$$

[1] Conventions differ in different communities. Some texts define Axiom 3 of a complex inner product to be linear in the *first slot* (not the second), which means the "standard" inner product on $\mathbb{C}^n$ must be changed to $\langle x, y \rangle = y^H x$ instead of $\langle x, y \rangle = x^H y$.

3. Linearity in the second slot: For all $x, y, z \in V$ and all scalars $a, b \in \mathbb{C}$ we have

$$\langle z, ax + by \rangle = a \langle z, x \rangle + b \langle z, y \rangle$$

The main difference between the three axioms of a real inner product and a complex inner product is that the symmetry axiom is changed to conjugate symmetry. Unfortunately, complex inner products are no longer linear in the first slot. Instead, it holds that $\langle ax + by, z \rangle = \overline{a} \langle x, z \rangle + \overline{b} \langle y, z \rangle$ for all $x, y, z \in V$ and all scalars $a, b \in \mathbb{C}$.

A vector space over the field $\mathbb{C}$ that has a complex inner product is called a *complex inner product space*. Results that hold for real inner product spaces, including the theory of projections and orthogonality, typically extend to complex inner product spaces with minor adjustments.

The *inner product norm* on a complex inner product space V is defined

$$||x|| = \sqrt{\langle x, x \rangle} \quad \forall x \in V$$

It can be shown that the inner product norm is always real, nonnegative, and satisfies the three properties of a norm (nonnegativity, scaling, triangle inequality, see Section 8.3). We have

$$||x + y||^2 = ||x||^2 + ||y||^2 + 2Re\left(\langle x, y \rangle\right) \quad \forall x, y \in V$$

where $Re(z)$ denotes the real part of $z \in \mathbb{C}$.

Two vectors $x, y \in V$ are said to be *orthogonal* if $\langle x, y \rangle = 0$. If $\{x_1, ..., x_k\}$ are vectors in V that are pairwise orthogonal then

$$\left|\left|\sum_{j=1}^{k} x_j\right|\right|^2 = \sum_{j=1}^{k} \left||x_j|\right|^2$$

If $\{v_1, ..., v_k\}$ are vectors in a complex inner product space V, we define the *Gram matrix* as the $k \times k$ matrix of inner products:

$$G = \left(\left\langle v_i, v_j \right\rangle\right) \quad , \quad G^H = G$$

Unlike real inner product spaces, the Gram matrix is not necessarily symmetric. However, it is easy to see that the Gram matrix is always Hermitian: $G^H = G$. It can be shown that $\{v_1, ..., v_k\}$ are linearly independent if and only if their Gram matrix G is invertible. Vectors $\{v_1, ..., v_k\}$ are orthonormal if and only if $G = I$.

10.3.1. Gram-Schmidt

The Gram-Schmidt procedure is the same as before. In particular, every nontrivial and finite dimensional complex inner product space has an orthonormal basis.

10.3.2. Cauchy-Schwarz inequality

The Cauchy-Schwarz inequality can be shown to be the same as before: If V is a complex inner product space then

$$|\langle x, y\rangle| \leq ||x|| \cdot ||y|| \quad \forall x, y \in V$$

with equality if and only if either $y = 0$ or $x = ay$ for some $a \in \mathbb{C}$.

10.3.3. Projection theorem

Projection results for complex inner product spaces are the same as before: Suppose V is a complex inner product space. Let S be a subspace of V. Fix $y \in V$. Define distances by the inner product norm. Then

- If a closest point in S to y exists, it must be unique.
- If $\dim(S) < \infty$ then a closest point in S to y exists.
- For either case $\dim(S) < \infty$ or $\dim(S) = \infty$: A vector p is the closest point in S to y if and only if $p \in S$ and the following orthogonality conditions hold:

$$\langle y - p, s\rangle = 0 \quad \forall s \in S$$

10.4. Example: The complex inner product space $\mathbb{C}^n$

Consider the complex vector space of n-tuples of complex numbers:

$$\mathbb{C}^n = \{(z_1, ..., z_n) : z_j \in \mathbb{C} \quad \forall j \in \{1, ..., n\}\}$$

The *standard inner product* on this space is the complex inner product defined by

$$\langle x, y\rangle = x^H y \quad \forall x, y \in \mathbb{C}^n$$

As an exercise, the reader can find an example where $x^H y \neq y^H x$. The reader can also verify that the three axioms for a complex inner product hold. For example, $x^H x \geq 0$ for all $x \in \mathbb{C}^n$, with equality if and only if $x = 0$.

A complex-valued matrix $A \sim m \times n$ has n columns that are themselves vectors in $\mathbb{C}^m$. The theory of projections can be extended to this complex world: Suppose we want to project a vector $y \in \mathbb{C}^m$ onto the column space of some $m \times n$ matrix A that has complex

entries. This projection vector p^* is defined as a solution to the following problem:

$$\begin{aligned} \text{Minimize:} \quad & \|y - p\| \\ \text{Subject to:} \quad & p \in Col(A) \end{aligned}$$

As in the theory of projections for real inner product spaces, the projection p^* can be shown to exist and to be unique. The above problem is equivalent to finding an $x \in \mathbb{C}^n$ to minimize $\|y - Ax\|$. Let $x^* \in \mathbb{C}^n$ be a solution (such an x^* exists but may not be unique). Then $p^* = Ax^*$. Also, if some other minimizer $\tilde{x}$ of $\|y - Ax\|$ is found, we must have $p^* = A\tilde{x}$. So $\tilde{x}$ minimizes $\|y - Ax\|$ if and only if $\tilde{x} \in x^* + Null(A)$.

The formula for projecting y onto $Col(A)$ can be derived using orthogonality principles similar to those for real inner product spaces: If we want to find $x \in \mathbb{C}^n$ to minimize $\|y - Ax\|$ then

$$x^* = (A^H A)^{-1} A^H y$$

$$p^* = Ax^* = A(A^H A)^{-1} A^H y$$

assuming $A^H A$ is invertible. These formulas are the same as their real counterparts in (9.14)-(9.15), with the exception that $A^\top$ is changed to A^H. Note that $A^H A$ is the Gram matrix for the columns of A. It can be shown that $A^H A$ is invertible if and only if the columns of A are linearly independent.

10.5. Example: Complex inner product space of square integrable functions

Fix a, b as real numbers that satisfy $a < b$. Consider the complex vector space V of complex-valued functions $f : [a, b] \to \mathbb{C}$ that are *square integrable,* so

$$\int_a^b |f(t)|^2 dt < \infty$$

where $|f(t)|$ denotes the magnitude of the complex number $f(t)$. The *standard inner product* on this space is

$$\langle f, g \rangle = \int_a^b \overline{f(t)} g(t) dt \quad \forall f, g \in V$$

which is the integral of the complex conjugate of $f(t)$ multiplied by $g(t)$. This is a valid complex inner product under the following definition of equivalence:

$$f \equiv g \iff \int_a^b |f(t) - g(t)|^2 dt = 0$$

This ensures the first axiom of inner products holds:

$$\begin{aligned}\langle f, f\rangle = 0 &\iff \int_a^b \overline{f(t)} f(t) dt = 0 \\ &\iff \int_a^b |f(t)|^2 dt = 0 \\ &\iff \int_a^b |f(t) - 0|^2 dt = 0 \\ &\iff f \equiv 0\end{aligned}$$

To project a function $f(t) \in V$ onto $Span\{g_1(t), ..., g_k(t)\}$, where $g_j(t) \in V$ for $j \in \{1, ...k\}$, we find the coefficients $\theta_j \in \mathbb{C}$ for the projection function $p(t) = \sum_{j=1}^k \theta_j g_j(t)$. The projection function satisfies the orthogonality conditions:

$$\langle f - p, g_i\rangle = 0 \quad \forall i \in \{1, ..., k\}$$

Since the inner product is no longer linear in the first slot, it is convenient to write these orthogonality conditions in the following equivalent form:

$$\langle g_i, f - p\rangle = 0 \quad \forall i \in \{1, ..., k\}$$

So for each $i \in \{1, ..., k\}$

$$\langle g_i, f\rangle = \langle g_i, p\rangle = \left\langle g_i, \sum_{j=1}^k \theta_j g_j \right\rangle = \sum_{j=1}^k \theta_j \left\langle g_i, g_j \right\rangle$$

That is

$$\underbrace{\begin{bmatrix} \langle g_1, f\rangle \\ \langle g_2, f\rangle \\ \vdots \\ \langle g_k, f\rangle \end{bmatrix}}_{h} = \underbrace{\begin{bmatrix} \langle g_1, g_1\rangle & \langle g_1, g_2\rangle & \cdots & \langle g_1, g_k\rangle \\ \langle g_2, g_1\rangle & \langle g_2, g_2\rangle & \cdots & \langle g_2, g_k\rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle g_k, g_1\rangle & \langle g_k, g_2\rangle & \cdots & \langle g_k, g_k\rangle \end{bmatrix}}_{G} \underbrace{\begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_k \end{bmatrix}}_{\theta}$$

so $h = G\theta$, where G is the $k \times k$ (complex-valued) Gram matrix for $\{g_1, ..., g_k\}$ and $h = (h_1, ..., h_k) \in \mathbb{C}^k$ is the vector with components $h_i = \langle g_i, f\rangle$. Since the projection exists, there is always a $\theta \in \mathbb{C}^k$ that solves $h = G\theta$. Also, any such solution θ yields the unique projection $p(t) = \sum_{j=1}^k \theta_j g_j(t)$. If $\{g_1, ..., g_k\}$ are linearly independent then G is invertible and there is one unique θ solution, namely, $\theta = G^{-1}h$.

If $\{g_1(t), ..., g_k(t)\}$ are orthonormal then $G = I$ and $\theta = h$, so

$$\theta_i = \langle g_i, f\rangle \quad \forall i \in \{1, ..., k\} \tag{10.3}$$

10.6. Fourier series via complex exponentials

Fix $T > 0$. Let V be the complex vector space of functions $f : [-T/2, T/2] \to \mathbb{C}$ that are square integrable. For each $n \in \mathbb{N}$, the following $2n+1$ functions are orthonormal in V:

$$\left\{ \frac{1}{\sqrt{T}} e^{2\pi i k t/T} \right\}_{k=-n}^{n} \tag{10.4}$$

since, with $\mathbb{Z}$ being the set of integers,

$$(m, k \in \mathbb{Z}, m \neq k) \implies \left\langle e^{2\pi i m t/T}, e^{2\pi i k t/T} \right\rangle = \int_{-T/2}^{T/2} e^{-2\pi i m t/T} e^{2\pi i k t/T} dt = 0$$

$$(k \in \mathbb{Z}) \implies \left\langle \frac{1}{\sqrt{T}} e^{2\pi i k t/T}, \frac{1}{\sqrt{T}} e^{2\pi i k t/T} \right\rangle = \frac{1}{T} \int_{-T/2}^{T/2} \underbrace{e^{-2\pi i k t/T} e^{2\pi i k t/T}}_{1} dt = 1$$

As time t increases, the functions $e^{2\pi i k t/T}$ spiral counter-clockwise in the unit circle of the complex plane when $k > 0$ (positive frequencies), and clockwise when $k < 0$ (negative frequencies).

Define S_n as the span of these $2n+1$ functions:

$$S_n = Span \left\{ \frac{1}{\sqrt{T}} e^{2\pi i k t/T} \right\}_{k=-n}^{n} \tag{10.5}$$

The projection of a function $f \in V$ onto S_n is

$$p_n(t) = \sum_{k=-n}^{n} c_k \frac{1}{\sqrt{T}} e^{2\pi i k t/T} \tag{10.6}$$

for some complex-valued coefficients $c_{-n}, ..., c_0, c_1, ..., c_n$. From (10.3) we have

$$c_k = \left\langle \frac{1}{\sqrt{T}} e^{2\pi i k t/T}, f(t) \right\rangle = \frac{1}{\sqrt{T}} \int_{-T/2}^{T/2} e^{-2\pi i k t/T} f(t) dt \quad \forall k \in \{-n, ..., n\} \tag{10.7}$$

The estimation of f gets better and better as n increases: The *Riesz-Fischer theorem* ensures that if $f \in V$ then, letting p_n be the projection of f onto S_n, we have

$$\lim_{n\to\infty} ||f - p_n||^2 = \lim_{n\to\infty} \int_{-T/2}^{T/2} |f(t) - p_n(t)|^2 dt = 0 \tag{10.8}$$

THEOREM 10.1. *(Parseval's identity for complex exponentials) If $f \in V$ then $||f||^2 < \infty$, where $||\cdot||$ uses the standard inner product norm, and*

$$||f||^2 = ||f - p_n||^2 + ||p_n||^2 \quad \forall n \in \mathbb{N}$$

where p_n is the projection of f onto the subspace S_n in (10.5)*. Also*

$$\|f\|^2 = \sum_{k=-\infty}^{\infty} |c_k|^2$$

where the complex scalars c_k are defined in (10.7)*, and $|c_k|$ denotes the magnitude of c_k.*

PROOF. By definition of f being in V, we know $\|f\|^2 < \infty$. Then for each $n \in \mathbb{N}$

$$\begin{aligned}\|f\|^2 &= \|f - p_n + p_n\|^2 \\ &= \|f - p_n\|^2 + \|p_n\|^2 + \underbrace{\langle f - p_n, p_n\rangle}_{0} + \underbrace{\langle p_n, f - p_n\rangle}_{0} \\ &= \|f - p_n\|^2 + \|p_n\|^2\end{aligned}$$

which uses the fact that, since p_n is the orthogonal projection of f onto the subspace S_n, we know $f - p_n$ is orthogonal to every function in S_n, including $p_n \in S_n$. Taking $n \to \infty$ and using (10.8) gives

$$\|f\|^2 = \lim_{n\to\infty} \|p_n\|^2$$

On the other hand, orthonormality of the complex exponentials implies

$$\|p_n\|^2 = \left\| \sum_{k=-n}^{n} c_k \frac{e^{2\pi ikt/T}}{\sqrt{T}} \right\|^2 = \sum_{k=-n}^{n} |c_k|^2 \quad \forall n \in N$$

□

10.6.1. Any interval of size T

For fixed interval size $T > 0$, the functions in (10.4) are also orthonormal when viewed as functions over any interval of size T. Specifically, fix $a \in \mathbb{R}$ and define V as the complex inner product space of square integrable functions $f : [a, a+T] \to \mathbb{C}$, with inner product

$$\langle f, g\rangle = \int_a^{a+T} \overline{f(t)} g(t) dt \quad \forall f, g \in V$$

Then the functions (10.4) are orthonormal and any $f \in V$ has the projection onto the span of these functions given by $p_n(t) = \sum_{k=-n}^{n} c_k \frac{e^{2\pi ikt/T}}{\sqrt{T}}$ with

$$c_k = \left\langle \frac{e^{2\pi ikt/T}}{\sqrt{T}}, f(t) \right\rangle \quad \forall k$$

While $p_n(t)$ is defined over $t \in [a, a+T]$, if it is extended to all $t \in \mathbb{R}$ it repeats T-periodically. This is because its complex exponentials are T-periodic functions over all $t \in \mathbb{R}$.

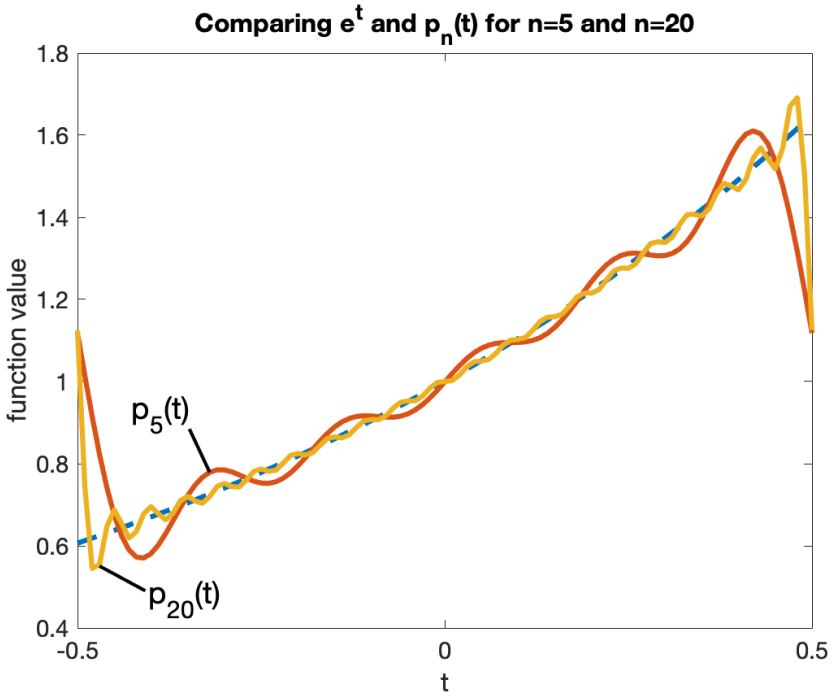


FIGURE 10.1. A comparison of e^t with its projection $p_n(t)$ for $n = 5$ and $n = 20$. The blue dashed curve is e^t.

10.6.2. Example with real $f(t)$

Consider $T = 1$ and $f(t) = e^t$ (a purely real-valued function). Then by (10.7)

$$c_k = \int_{-1/2}^{1/2} e^{-2\pi i k t} e^t dt = \frac{2\sinh(1/2 - i\pi k)}{1 - 2i\pi k} \quad \forall k \in \mathbb{Z}$$

and the projection of e^t onto S_n is given by $p_n(t)$ in (10.6) (see Fig. 10.1). In general, when $f(t)$ is purely real then $p_n(t)$ will be purely real ($c_0 \in \mathbb{R}$, and for $k \in \mathbb{N}$ the imaginary parts of the contribution of frequencies k and $-k$ cancel). In fact, when $f(t)$ is purely real, the $p_n(t)$ function is identical to the Fourier representation with $\cos(\cdot)$ and $\sin(\cdot)$ functions described in Chapter 9.

10.7. Are dimensions doubled?

Each vector $z \in \mathbb{C}^n$ has the form $z = x + iy$ for some $x, y \in \mathbb{R}^n$. Thus, it takes $2n$ real numbers to specify z. So does $\mathbb{C}^n$ have dimension n, or dimension $2n$? Specifically, define

$$e_1 = (1, 0, 0, 0, \ldots, 0) \quad , \quad e_2 = (0, 1, 0, 0, \ldots, 0) \quad , \quad e_3 = (0, 0, 1, 0, \ldots, 0)$$

and so on. Which one of the following is a basis for $\mathbb{C}^n$?

- Choice 1: $\{e_1, e_2, ..., e_n\}$.
- Choice 2: $\{e_1, e_2, ..., e_n, ie_1, ie_2, ..., ie_n\}$.

The answer is choice 1: $\{e_1, ..., e_n\}$ is still the standard basis for $\mathbb{C}^n$. So $\dim(\mathbb{C}^n) = n$. Intuitively, this is because each scalar $\theta \in \mathbb{C}$ is a 2-parameter object $\theta = a + bi$ for $a, b \in \mathbb{R}$.

PART IV.

Eigenvalues, singular values, and their applications

CHAPTER 11.

Determinants

This chapter describes algebraic properties of the determinant of a matrix. It considers a general field $\mathbb{F}$. Section 11.12 considers geometric properties when $\mathbb{F} = \mathbb{R}$.

Let n be a positive integer. Define

$$\mathbb{F}^{n\times n} = \{n \times n \text{ matrices with entries in } \mathbb{F}\}$$

Given a matrix $A \in \mathbb{F}^{n\times n}$, we use $row_i(A)$ to denote the row vector formed by row i of matrix A. For example

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

$$row_1(A) = \begin{bmatrix} 0 & 1 & 1 \end{bmatrix}$$
$$row_2(A) = \begin{bmatrix} 3 & 0 & 2 \end{bmatrix}$$
$$row_3(A) = \begin{bmatrix} 2 & 1 & 0 \end{bmatrix}$$

11.1. Axioms of determinants

There is a *determinant function* $\det : \mathbb{F}^{n\times n} \to \mathbb{F}$ that maps each matrix $A \in \mathbb{F}^{n\times n}$ to a scalar $\det(A) \in \mathbb{F}$. The determinant function satisfies the following three axioms:

1. Axiom 1: $\det(I) = 1$ (where I is the $n \times n$ identity matrix). For example:

$$\det\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) = 1$$

2. Axiom 2: If A has a repeated row (so $row_i(A) = row_j(A)$ for some $i \neq j$) then $\det(A) = 0$. For example

$$\det\left(\begin{bmatrix} 1 & 1 & 2 \\ 0 & 1 & 0 \\ 1 & 1 & 2 \end{bmatrix}\right) = 0 \quad , \quad \det\left(\begin{bmatrix} 3 & 3 & 1 \\ -5 & 3 & 7 \\ -5 & 3 & 7 \end{bmatrix}\right) = 0$$

3. Axiom 3: The determinant is linear in every row of a matrix. Specifically, given any row $r \in \{1, \ldots, n\}$ and any scalars $\alpha, \beta \in \mathbb{F}$, we have

$$\det(C) = \alpha \det(A) + \beta \det(B)$$

whenever A, B, C are matrices in $\mathbb{F}^{n \times n}$ that have identical rows i for all $i \neq r$, and that satisfy

$$row_r(C) = \alpha row_r(A) + \beta row_r(B)$$

For example

$$\det\left(\begin{bmatrix} 1 & 0 & 9 \\ \alpha a_1 + \beta b_1 & \alpha a_2 + \beta b_2 & \alpha a_3 + \beta b_3 \\ 3 & 9 & -2 \end{bmatrix}\right)$$
$$= \alpha \det\left(\begin{bmatrix} 1 & 0 & 9 \\ a_1 & a_2 & a_3 \\ 3 & 9 & -2 \end{bmatrix}\right) + \beta \det\left(\begin{bmatrix} 1 & 0 & 9 \\ b_1 & b_2 & b_3 \\ 3 & 9 & -2 \end{bmatrix}\right)$$

A special case of linearity in each row is that a scalar can be pulled out of any row:

$$\det\left(\begin{bmatrix} 1 & 0 & 9 \\ a & b & c \\ \theta & -4\theta & -2\theta \end{bmatrix}\right) = \theta \det\left(\begin{bmatrix} 1 & 0 & 9 \\ a & b & c \\ 1 & -4 & -2 \end{bmatrix}\right) \quad \forall \theta \in \mathbb{F}$$

11.2. Existence and uniqueness

It is not obvious that there exists a function that satisfies the three axioms above. It can be proven that for any field $\mathbb{F}$ and any $n \in \mathbb{N}$, such a determinant function $\det : \mathbb{F}^{n\times n} \to \mathbb{F}$ exists. The main sections of this chapter assume existence and then develop properties of the determinant to show it is *unique*. That is, there is one and only one function $f : \mathbb{F}^{n\times n} \to \mathbb{F}$ that satisfies the three axioms, and we call that function $f(A) = \det(A)$. The question of existence is treated later (Section 11.11).

To help understand the issues of existence and uniqueness, here we construct two distinct functions that satisfy only the first two axioms. Assume $n \geq 2$. Define $f : \mathbb{F}^{n\times n} \to \mathbb{F}$ and $g : \mathbb{F}^{n\times n} \to \mathbb{F}$ by

$$f(A) = \begin{cases} 1 & \text{if } A = I \\ 0 & \text{if } A \text{ has a repeated row} \\ 0 & \text{else} \end{cases}$$

$$g(A) = \begin{cases} 1 & \text{if } A = I \\ 0 & \text{if } A \text{ has a repeated row} \\ 1 & \text{else} \end{cases}$$

The functions f and g are well defined because if $A = I$ then A does not have a repeated row. Both functions map each matrix in $\mathbb{F}^{n\times n}$ to either 0 or 1, and 0 and 1 are indeed scalars in $\mathbb{F}$. The functions f and g are *distinct* because if we let P be the $n \times n$ permutation matrix formed by swapping rows 1 and 2 of I then

$$f(P) = 0 \quad , \quad g(P) = 1$$

It is easy to see that both f and g satisfy the first two axioms of determinants. Thus, those two axioms are inclusive enough to allow *existence* (the axioms are not contradictory). However, they are not stringent enough to ensure *uniqueness* (Axiom 3 provides this). Neither f nor g is the true determinant function because neither satisfies Axiom 3.

11.3. Basic consequences

Assume throughout that $\det(\cdot)$ satisfies the three axioms of determinants.

LEMMA 11.1. *If a square matrix A has a row of all zeros, then* $\det(A) = 0$.

PROOF. The scalar $0 \in \mathbb{F}$ can be pulled out of the all-zero row. For example:

$$\det\left(\begin{bmatrix} a & b & c \\ 0 & 0 & 0 \\ d & e & f \end{bmatrix}\right) = \det\left(\begin{bmatrix} a & b & c \\ 0 \cdot 0 & 0 \cdot 0 & 0 \cdot 0 \\ d & e & f \end{bmatrix}\right) = 0 \cdot \det\left(\begin{bmatrix} a & b & c \\ 0 & 0 & 0 \\ d & e & f \end{bmatrix}\right)$$

□

LEMMA 11.2. *The determinant of a purely diagonal square matrix is the product of its diagonal elements.*

PROOF. The diagonal entries can be pulled out row-by-row. For example:

$$\det\left(\begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix}\right) = abc \cdot \det\left(\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\right) = abc$$

□

LEMMA 11.3. *Multiplying one row of a square matrix by a scalar θ and adding the result into another row does not change the determinant.*

PROOF. This uses linearity in each row (Axiom 3). For simplicity we demonstrate the proof for a 3×3 matrix where we multiply row 2 by θ and add the result into row 3:

$$\begin{aligned}
\det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g + \theta d & h + \theta e & i + \theta f \end{bmatrix}\right) &= \det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ 1 \cdot g + \theta d & 1 \cdot h + \theta e & 1 \cdot i + \theta f \end{bmatrix}\right) \\
&= 1 \cdot \det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right) + \theta \cdot \underbrace{\det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ d & e & f \end{bmatrix}\right)}_{0} \\
&= \det\left(\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}\right)
\end{aligned}$$

where the term with the underbrace is 0 because it has a repeated row. □

LEMMA 11.4. *Swapping two rows of a square matrix multiplies the determinant by* –1.

PROOF. The previous lemma shows that performing an important row operation on a matrix does not change the determinant. We can swap rows by performing a sequence of these row operations. For simplicity we demonstrate the idea on a 3×3 matrix and swap rows 2 and 3. Define:

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} \quad , \quad \tilde{A} = \begin{bmatrix} a & b & c \\ g & h & i \\ d & e & f \end{bmatrix}$$

We have

$$\underbrace{\begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}}_{A} \overset{(a)}{\to} \begin{bmatrix} a & b & c \\ d & e & f \\ g-d & h-e & i-f \end{bmatrix} \overset{(b)}{\to} \begin{bmatrix} a & b & c \\ g & h & i \\ g-d & h-e & i-f \end{bmatrix} \overset{(c)}{\to} \underbrace{\begin{bmatrix} a & b & c \\ g & h & i \\ -d & -e & -f \end{bmatrix}}_{B}$$

where (a) multiplies row 2 by -1 and adds into row 3; (b) adds row 3 into row 2; (c) multiplies row 2 by -1 and adds into row 3. The previous lemma ensures the determinant is unchanged by these operations. Define B as the matrix in the final underbrace. Then $\det(A) = \det(B)$ and so

$$\det(A) = \det(B) = -\det(\tilde{A})$$

where the final equality uses linearity by pulling the scalar –1 out of row 3 of B. □

Caveat: Our three axioms of determinants are useful for matrices in a *general field* $\mathbb{F}$. Lemma 11.4 shows these axioms imply that swapping any two rows multiplies the determinant by –1. In other texts, such as Strang (2023), determinants are treated in the special case $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$. In that case, Axiom 2 can be equivalently replaced by an axiom, call it Axiom 2', that simply *assumes* that swapping any two rows multiplies the determinant by –1. For the fields $\mathbb{R}$ or $\mathbb{C}$, Axiom 2' can be used to immediately prove the determinant of a matrix with a repeated row is zero (so the missing Axiom 2 is recovered). Indeed, if A has a repeated row then swapping those two identical rows yields A again, so Axiom 2' implies

$$\det(A) = -det(A)$$

which, in any field for which $-1 \neq 1$, we obtain $\det(A) = 0$. Of course, the equation

$\det(A) = -\det(A)$ does *not* imply $\det(A) = 0$ if we are using the binary field $\mathbb{F}_2 = \{0, 1\}$.

11.4. Elementary row operations

Recall that Gaussian elimination uses three types of *elementary row operations*. Each type corresponds to left-multiplication by an *elementary matrix* of that type.

1. Type 1: Swap two rows of a matrix:

$$E_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \implies E_2 \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} a & b & c \\ g & h & i \\ d & e & f \end{bmatrix}$$

   Note that a type-1 elementary matrix E_2 can be obtained by swapping the same two rows of the identity matrix of the same size, so Lemma 11.4 implies

$$\det(E_2) = -\det(I) = -1 \neq 0$$

   Since Lemma 11.4 ensures that swapping two rows of a matrix multiplies the determinant by -1, we have for any type-1 elementary matrix and any matrix A of the same size:

$$\det(E_2 A) = -det(A) = \det(E_2)\det(A)$$

   Also, it is easy to see that all type-1 elementary matrices are invertible.

2. Type 2: Multiply row i of a matrix by a scalar $\theta \neq 0$:

$$E_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \theta & 0 \\ 0 & 0 & 1 \end{bmatrix} \implies E_1 \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} a & b & c \\ \theta d & \theta e & \theta f \\ g & h & i \end{bmatrix}$$

   Note that a type-2 elementary matrix E_1 can be obtained by performing that row operation on the identity matrix I of the same size as E_1. So $E_1 = E_1 I$. Since $\theta \neq 0$, it is easy to see that every type-2 elementary matrix is invertible. Since every type-2 elementary matrix is purely diagonal with θ on a single entry and 1 on all other entries, Lemma 11.2 ensures

$$\det(E_1) = \theta \neq 0$$

   Further, since multiplying a row of a matrix by θ scales the determinant by that amount (by linearity of determinants in that row), we have for any type-2 elementary matrix

and any square matrix A of the same size

$$\det(E_1 A) = \theta \det(A) = \det(E_1)\det(A)$$

3. Type 3: Multiply row i by a scalar θ and add the result into another row $j \neq i$:

$$E_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \theta & 0 & 1 \end{bmatrix} \implies E_3 \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix} = \begin{bmatrix} a & b & c \\ d & e & f \\ g+\theta a & h+\theta b & i+\theta c \end{bmatrix}$$

Lemma 11.3 ensures a type-3 row operation does not change the determinant of a matrix A. So for any type-3 elementary matrix E_3 and any matrix A of the same size we have

$$\det(E_3 A) = \det(A)$$

Since $E_3 = E_3 I$ we obtain

$$\det(E_3) = \det(E_3 I) = \det(I) = 1 \neq 0$$

and furthermore

$$\det(E_3 A) = \det(A) = 1 \cdot \det(A) = \det(E_3)\det(A)$$

Also, it is easy to see that all type-3 matrices are invertible.

In summary, if E is an elementary matrix then it is one of three types. Regardless of the type, E must be invertible, E must have nonzero determinant, and for any matrix A of the same size we must have

$$\boxed{\det(EA) = \det(E)\det(A)}$$

Furthermore,

$$\det(E) = \begin{cases} \theta & \text{if } E \text{ is type-2 (with parameter } \theta \neq 0) \\ -1 & \text{if } E \text{ is type-1} \\ 1 & \text{if } E \text{ is type-3 (with parameter } \theta) \end{cases}$$

This means that determinants can be computed by Gaussian elimination: If a square matrix A can be row-reduced to another matrix U by a sequence of elementary matrices,

so that

$$E_k \cdots E_2 E_1 A = U$$

then we immediately obtain

$$\det(E_k)\det(E_{k-1})\cdots\det(E_1)\det(A) = \det(U)$$

and since elementary matrices have nonzero determinants, we obtain

$$\det(A) = \det(E_k)^{-1}\det(E_{k-1})^{-1}\cdots\det(E_1)^{-1}\det(U)$$

Thus, if U has an easy-to-compute determinant, this allows a simple way to compute $\det(A)$.

11.5. Deeper consequences

Assume all matrices in the lemmas below are square and have entries in some field $\mathbb{F}$.

LEMMA 11.5. *A is invertible if and only if* $\det(A) \neq 0$.

PROOF. Suppose A is invertible. Then it can be row-reduced to the identity matrix I by a sequence of elementary row operations:

$$E_k E_{k-1} \cdots E_1 A = I$$

and so

$$\det(E_k)\det(E_{k-1})\cdots\det(E_1)\det(A) = \det(I) = 1$$

It follows that $\det(A) \neq 0$.

Now suppose A is not invertible. Then it can be row-reduced to a matrix U with an all-zero row:

$$E_k E_{k-1} \cdots E_1 A = U$$

and so

$$\det(E_k)\det(E_{k-1})\cdots\det(E_1)\det(A) = \det(U) = 0$$

where $\det(U) = 0$ because U has an all-zero row (recall Lemma 11.1). Since $\det(E_i) \neq 0$ for all $i \in \{1, \ldots, k\}$, we must have $\det(A) = 0$. □

LEMMA 11.6. *If A and B are two square matrices of the same size then*

$$\det(AB) = \det(A)\det(B)$$

Likewise, if $A_1, \ldots, A_k$ are square matrices of the same size then

$$\det(A_1 \cdots A_k) = \det(A_1) \cdots \det(A_k)$$

PROOF. If A is invertible then $A = E_k \cdots E_1$ for some positive integer k and some elementary matrices E_i, so

$$\det(AB) = \det(E_k \cdots E_1 B) = \det(E_k) \cdots \det(E_1)\det(B) = \det(E_k \cdots E_1)\det(B) = \det(A)\det(B)$$

On the other hand, if A is not invertible then AB is not invertible and $\det(A) = 0$ and $\det(AB) = 0$. □

LEMMA 11.7. *(Uniqueness) Fix $\mathbb{F}$ as a field. If* $\det(A)$ *is a function that maps a matrix $A \in \mathbb{F}^{n \times n}$ to a scalar* $\det(A) \in \mathbb{F}$ *that satisfies Axioms 1, 2, 3 for determinants, then it is the* unique *such function.*

PROOF. Assume the $\det(\cdot)$ function satisfies the three axioms of determinants. Fix $A \in \mathbb{F}^{n \times n}$. We show there is one and only one possible value of $\det(A)$. If A is not invertible, the previous lemma shows that $\det(A) = 0$. On the other hand, if A is invertible then it can be row-reduced to the identity matrix by some particular elementary matrices, so

$$E_k E_{k-1} \cdots E_1 A = I$$

and therefore

$$\det(E_k)\det(E_{k-1}) \cdots \det(E_1)\det(A) = \det(I) = 1$$

and so

$$\det(A) = \det(E_k)^{-1}\det(E_{k-1})^{-1} \cdots \det(E_1)^{-1}$$

□

LEMMA 11.8. *(Upper triangular matrix) If U is a square upper-triangular matrix then* $\det(U)$

is the product of the diagonal elements of U. For example

$$\det\left(\begin{bmatrix} a & b & c \\ 0 & d & e \\ 0 & 0 & f \end{bmatrix}\right) = adf$$

PROOF. If U has at least one diagonal entry that is zero then it is not invertible and so its determinant is zero (which is indeed the product of its diagonal entries). If U has all nonzero diagonal entries then it can be row-reduced to a purely diagonal matrix (with those same diagonal entries) using only type-3 operations, which do not change the determinant. The determinant of that purely diagonal matrix is the product of its diagonal entries (by Lemma 11.2). □

LEMMA 11.9. *If A is a square matrix then* $\det(A) = \det(A^\top)$.

PROOF. Recall that A is invertible if and only if $A^\top$ is invertible. If A is not invertible then $\det(A) = 0$ and so $\det(A^\top) = 0$. It remains to consider the case when A is invertible. In this case, A is the product of elementary matrices:

$$\begin{aligned} A = E_1 E_2 \cdots E_k &\implies \det(A) = \det(E_1)\det(E_2)\cdots\det(E_k) \\ A^\top = E_k^\top E_{k-1}^\top \cdots E_2^\top E_1^\top &\implies \det(A^\top) = \det(E_k^\top)\det(E_{k-1}^\top)\cdots\det(E_2^\top)\det(E_1^\top) \end{aligned}$$

and it is left as an exercise to show that $\det(E) = \det(E^\top)$ for any elementary matrix E (hint: consider the three types separately). □

LEMMA 11.10. *The determinant of a lower-triangular matrix is the product of its diagonal entries.*

PROOF. The transpose of a lower triangular matrix is an upper triangular matrix. □

LEMMA 11.11. *The determinant is a linear function of every column. Also, swapping any two columns multiplies the determinant by* -1.

PROOF. Axiom 3 implies the determinant is a linear function of each row of A. We also know that swapping any two rows multiplies the determinant by -1. The result follows immediately by the fact that the determinant is unchanged by the transpose operation. □

11.6. Determinants by Gaussian elimination

We know the determinant of an upper-triangular matrix is the product of its diagonal entries. One of the best ways to compute $\det(A)$ is to perform Gaussian elimination to row-reduce A to an upper triangular matrix U

$$E_k \cdots E_1 A = U \implies \det(E_k) \cdots \det(E_1) \det(A) = \det(U)$$

Since the determinant of a type-3 elementary matrix is 1, and the determinant of a type-1 elementary matrix is -1, we obtain

$$(-1)^{\text{Num row swaps}} \cdot (\text{Product of } \theta_i \text{ values of each type-2}) \cdot \det(A) = \det(U)$$

11.7. Determinant with block upper triangular structure

Fix $\mathbb{F}$ as a field. Suppose k, m are positive integers, B is a square $k \times k$ matrix, D is a square $m \times m$ matrix, and we have a square matrix A of the following "block upper triangular" form

$$A = \begin{bmatrix} B & C \\ 0 & D \end{bmatrix}$$

where C is some $k \times m$ matrix; 0 represents the all-zero matrix of size $m \times k$. Thus, A is $n \times n$ where $n = k + m$. It can be shown that

$$\det(A) = \det(B) \det(D)$$

A similar result holds when A has a lower triangular block structure.

The proof is as follows: If either B or D is *not* invertible, then it can be shown that A has a nontrivial null space (exercise), so $\det(A) = 0$ and $\det(B) \det(D) = 0$, so the result is true. Now suppose both B and D are invertible. Then they can be row reduced to the identity matrix by a sequence of elementary matrices:

$$E_r \cdots E_1 B = I \text{ (of size } k \times k) \tag{11.1}$$

$$F_s \cdots F_1 D = I \text{ (of size } m \times m) \tag{11.2}$$

for some positive integers r, s. Thus

$$\left(\underbrace{\begin{bmatrix} E_r & 0 \\ 0 & I \end{bmatrix}}_{\tilde{E}_r} \cdots \underbrace{\begin{bmatrix} E_1 & 0 \\ 0 & I \end{bmatrix}}_{\tilde{E}_1}\right)\left(\underbrace{\begin{bmatrix} I & 0 \\ 0 & F_s \end{bmatrix}}_{\tilde{F}_s} \cdots \underbrace{\begin{bmatrix} I & 0 \\ 0 & F_1 \end{bmatrix}}_{\tilde{F}_1}\right)\underbrace{\begin{bmatrix} B & C \\ 0 & D \end{bmatrix}}_{A} = \underbrace{\begin{bmatrix} I & C \\ 0 & I \end{bmatrix}}_{\det=1}$$

The matrix on the right-hand-side has determinant 1 because it is upper triangular with all diagonals 1. Then

$$\left(\det(\tilde{E}_r)\cdots\det(\tilde{E}_1)\right)\cdot\left(\det(\tilde{F}_s)\cdots\det(\tilde{F}_1)\right)\cdot\det(A) = 1$$

Notice that each matrix $\tilde{E}_i$ is elementary with the same type and same determinant as E_i. Similar for $\tilde{F}_j$. Thus

$$\begin{aligned}\det(\tilde{E}_i) &= \det(E_i) \quad \forall i \in \{1,\ldots,r\}\\ \det(\tilde{F}_j) &= \det(F_j) \quad \forall j \in \{1,\ldots,s\}\end{aligned}$$

Thus

$$(\det(E_r)\cdots\det(E_1))\cdot(\det(F_s)\cdots\det(F_1))\cdot\det(A) = 1$$

The result follows by observing (11.1)-(11.2) imply

$$\det(E_r)\cdots\det(E_1) = 1/\det(B) \quad , \quad \det(F_s)\cdots\det(F_1) = 1/\det(D)$$

11.8. Determinant of 2×2

Fix $\mathbb{F}$ as a field and let

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

Observe that

$$\begin{bmatrix} c & d \end{bmatrix} = \begin{bmatrix} c & 0 \end{bmatrix} + \begin{bmatrix} 0 & d \end{bmatrix}$$

Since the determinant is linear in row 2 we obtain

$$\det(A) = \det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) \overset{(a)}{=} \det\left(\begin{bmatrix} a & b \\ c & 0 \end{bmatrix}\right) + \det\left(\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}\right)$$

$$\overset{(b)}{=} \det\left(\begin{bmatrix} c & 0 \\ a & b \end{bmatrix}\right) + \det\left(\begin{bmatrix} a & b \\ 0 & d \end{bmatrix}\right)$$
$$\overset{(c)}{=} -bc + ad$$

where (a) uses linearity in row 2; (b) uses the fact that swapping the rows multiplies the determinant by –1; (c) observes that we have a lower-triangular and an upper-triangular matrix. This yields the famous formula

$$\boxed{\det\left(\begin{bmatrix} a & b \\ c & d \end{bmatrix}\right) = ad - bc}$$

The 2×2 matrix A is invertible if and only if $ad \neq bc$. It is easy to verify that if A is invertible then

$$\boxed{A^{-1} = (ad - bc)^{-1} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}}$$

where $(ad - bc)^{-1}$ is the multiplicative inverse of the nonzero scalar $ad - bc$ in the field $\mathbb{F}$ of interest. This can be shown by multiplying the above matrix by A to obtain the 2×2 identity matrix.

11.9. Determinant of 3×3 and $n \times n$

Consider the following 3×3 matrix:

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

For the first row, we have

$$\begin{bmatrix} a & b & c \end{bmatrix} = a\begin{bmatrix} 1 & 0 & 0 \end{bmatrix} + b\begin{bmatrix} 0 & 1 & 0 \end{bmatrix} + c\begin{bmatrix} 0 & 0 & 1 \end{bmatrix}$$

Then

$$\det(A) \overset{(a)}{=} a \det\left(\begin{bmatrix} 1 & 0 & 0 \\ d & e & f \\ g & h & i \end{bmatrix}\right) + b \det\left(\begin{bmatrix} 0 & 1 & 0 \\ d & e & f \\ g & h & i \end{bmatrix}\right) + c \det\left(\begin{bmatrix} 0 & 0 & 1 \\ d & e & f \\ g & h & i \end{bmatrix}\right) \tag{11.3}$$

$$\overset{(b)}{=} a\det\left(\begin{bmatrix} 1 & 0 & 0 \\ d & e & f \\ g & h & i \end{bmatrix}\right) - b\det\left(\begin{bmatrix} 1 & 0 & 0 \\ e & d & f \\ h & g & i \end{bmatrix}\right) + c\det\left(\begin{bmatrix} 1 & 0 & 0 \\ f & d & e \\ i & g & h \end{bmatrix}\right) \tag{11.4}$$

$$\overset{(c)}{=} a\det\left(\begin{bmatrix} e & f \\ h & i \end{bmatrix}\right) - b\det\left(\begin{bmatrix} d & f \\ g & i \end{bmatrix}\right) + c\det\left(\begin{bmatrix} d & e \\ g & h \end{bmatrix}\right) \tag{11.5}$$

where (a) holds by linearity in the first row; (b) holds because: For the second matrix we have moved the second column all the way to the left by a single column swap, for the third matrix we have moved the third column all the way to the left by two column swaps; (c) holds because of the lower block triangular structure (and the fact that the determinant of the 1×1 identity matrix is 1).

By the same reasoning, the determinant of a $n \times n$ matrix $A = (a_{ij})$ can be computed similarly. Using linearity in the first row and the fact that it takes an odd number of column swaps to move an even-indexed column all the way to the left, and an even number of column swaps to move an odd-indexed column all the way to the left, we have

$$\det(A) = \sum_{j=1}^{n} a_{1j}(-1)^{1+j}\det(A_{1j}) \tag{11.6}$$

where A_{1j} is the $(n-1) \times (n-1)$ matrix formed by removing row 1 and column j from matrix A.

Linearity holds in any row i (not just row 1), so the same technique can be done using linearity in row i and the fact that it takes $i + j - 2$ row or column swaps to move the (i, j) entry all the way to the upper left:

$$\det(A) = \sum_{j=1}^{n} a_{ij}(-1)^{i+j}\det(A_{ij}) \quad \forall i \in \{1, \ldots, n\} \tag{11.7}$$

where A_{ij} is the $(n-1) \times (n-1)$ matrix formed by removing row i and column j from matrix A. The same can be done using linearity in a given column j:

$$\det(A) = \sum_{i=1}^{n} a_{ij}(-1)^{i+j}\det(A_{ij}) \quad \forall j \in \{1, \ldots, n\} \tag{11.8}$$

The formulas (11.7) and (11.8) are called *Laplace expansion formulas*, also called *cofactor expansion formulas* (see next section). For simplicity of calculation, it is often a good idea to determine which row or column to use by using the one with the largest number of zeros.

11.10. Cofactor formulas for $\det(A)$ and A^{-1}

Let $\mathbb{F}$ be a field. Let A be a $n \times n$ matrix. For each pair (i, j) we define the *cofactor*

$$c_{ij} = (-1)^{i+j} \det(A_{ij}) \quad \forall i, j \in \{1, ..., n\}$$

With this definition, the cofactor expansion formulas (11.7)-(11.8) become

$$\det(A) = \sum_{j=1}^{n} a_{ij} c_{ij} \quad \forall i \in \{1, ..., n\} \tag{11.9}$$

$$\det(A) = \sum_{i=1}^{n} a_{ij} c_{ij} \quad \forall j \in \{1, ..., n\} \tag{11.10}$$

Let $C = (c_{ij})$ be the $n \times n$ matrix of cofactors, called the *cofactor matrix of A* (also written $\text{Cofactor}(A)$):

$$C = \text{Cofactor}(A) = (c_{ij})$$

Define the *adjugate* of matrix A, written $\text{adj}(A)$, as the transpose of C:

$$\text{adj}(A) = C^{\top}$$

Thus, C and $\text{adj}(A)$ are both $n \times n$ matrices. Also, the (i, j) entry of $\text{adj}(A)$ is the (j, i) entry of C:

$$(\text{adj}(A))_{ij} = c_{ji} \quad \forall i, j \in \{1, ..., n\} \tag{11.11}$$

11.10.1. Cofactor formula for A^{-1}

LEMMA 11.12. *Let $\mathbb{F}$ be a field. Let A be a $n \times n$ matrix and assume $\det(A) \neq 0$. Then*

$$A^{-1} = \frac{1}{\det(A)} adj(A)$$

PROOF. Define $B = \frac{1}{\det(A)} \text{adj}(A)$. By Theorem 5.4 it suffices to prove that $AB = I$. That is, we want to show for all $i, j \in \{1, ..., n\}$ that

$$(AB)_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

By definition of B together with (11.11), the (i, j) entry of B is

$$(B)_{ij} = \frac{c_{ji}}{\det(A)} \quad \forall i, j \in \{1, ..., n\} \tag{11.12}$$

Fix $i, j \in \{1, \ldots, n\}$. By the matrix multiplication definition (2.5) we have

$$\begin{aligned}(AB)_{ij} &= \sum_{u=1}^{n} a_{iu}(B)_{uj} \\ &= \frac{1}{\det(A)} \sum_{u=1}^{n} a_{iu} c_{ju}\end{aligned} \tag{11.13}$$

where the final equality uses (11.12).

- Case $i = j$: If $i = j$ the equation (11.13) becomes

$$\begin{aligned}(AB)_{ii} &= \frac{1}{\det(A)} \sum_{u=1}^{n} a_{iu} c_{iu} \\ &\overset{(a)}{=} \frac{1}{\det(A)} \det(A) = 1\end{aligned}$$

  where (a) uses equation (11.9).

- Case $i \neq j$. If $i \neq j$, define the $n \times n$ matrix D by replacing row j of A with row i of A. Since D has two repeated rows, we know $\det(D) = 0$. On the other hand, since rows i and j of D are identical to row i of A, we have

$$d_{iu} = d_{ju} = a_{iu} \quad \forall u \in \{1, \ldots, n\} \tag{11.14}$$

  Also, along row j, the cofactor matrix for D (which depends only on entries of D that are *not* on row j) is the same as that of A:

$$(\text{Cofactor}(D))_{ju} = c_{ju} \quad \forall u \in \{1, \ldots, n\} \tag{11.15}$$

  Thus, starting from (11.13), we have

$$\begin{aligned}(AB)_{ij} &= \frac{1}{\det(A)} \sum_{u=1}^{n} a_{iu} c_{ju} \\ &\overset{(a)}{=} \frac{1}{\det(A)} \sum_{u=1}^{n} d_{ju} c_{ju} \\ &\overset{(b)}{=} \frac{1}{\det(A)} \sum_{u=1}^{n} d_{ju} (\text{Cofactor}(D))_{ju} \\ &\overset{(c)}{=} \frac{1}{\det A} \det(D) = 0\end{aligned}$$

  where (a) holds by substituting (11.14); (b) holds by substituting (11.15); (c) holds by the cofactor expansion formula (11.9); the final equality uses $\det(D) = 0$.

□

11.10.2. Exponential complexity of the cofactor formula for determinants

The cofactor formula for $\det(A)$ is useful for computing determinants of 3×3 or 4×4 matrices. For large matrices, Gaussian elimination is preferred. Indeed, a computer can use Gaussian elimination to easily compute the determinant of a 70×70 matrix, while the cofactor formula would require more computational steps than the current estimate of the number of atoms in the universe. To see why, let $f(n)$ be the worst-case computational complexity of computing the determinant of a $n \times n$ matrix using the cofactor formula on row 1. Here we measure complexity by the number of *lookup-multiply-add* operations. Let $f(1) = 1$ (we look up the value of the only entry of a 1×1 matrix, and that is its determinant). So computing $\det(A_{1j})$ for a given $j \in \{1, \ldots, n\}$ requires $f(n-1)$ complexity. By formula (11.6), we do this n times, then add up the weighted sum of results, so

$$f(n) \geq nf(n-1) \quad \forall n \in \{2, 3, 4, \ldots\}$$

which means $f(n) \geq n!$ for all $n \in \{1, 2, 3, \ldots\}$. Thus

$$f(70) \geq 70! > 10^{85}$$

Of course, determinants are easy to calculate if the matrix is upper triangular or lower triangular. Also, the computation of a determinant can be greatly reduced if the matrix has a block upper triangular or block lower triangular form.

11.11. Existence of the determinant function

Up to now, we have assumed *existence* of a function $f : \mathbb{F}^{n \times n} \to \mathbb{F}$ that satisfies the three axioms of determinants. We know that if such a determinant function exists, it must be unique and must satisfy the various lemmas developed above. Here we prove existence for a general $n \times n$ matrix. Let S_n denote the set of all $n!$ permutations of the ordered list $(1, \ldots, n)$. Let $A = (a_{ij})$ be a $n \times n$ matrix. Define the function $f : \mathbb{F}^{n \times n} \to \mathbb{F}$ by

$$f(A) = \sum_{\sigma \in S_n} sign(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)} \tag{11.16}$$

where $\sigma(i)$ denotes the ith entry in the permuted list σ, and $sign(\sigma) = 1$ if the permutation σ is obtained from $(1, \ldots, n)$ by an even number of swaps, and -1 if it is obtained by an odd number of swaps. This function certainly exists. Further, it can be shown that it satisfies the three axioms of determinants (see Exercise 11.8). This proves existence. Further, by

uniqueness of the determinant, it must be that $f(A) = \det(A)$, that is

$$\boxed{\det(A) = \sum_{\sigma \in S_n} sign(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}}$$

This formula involves a sum over $n!$ terms, so it is often too complicated to use for computing $\det(A)$. However, it provides an immediate proof of the following.

LEMMA 11.13. *Consider the real field $\mathbb{F} = \mathbb{R}$. If all entries of a square matrix A are integers, then $\det(A)$ is also an integer.*

PROOF. We know $\det(A) = f(A)$ where $f(A)$ is given in (11.16). For each permutation σ, we have $sign(\sigma) \in \{1, -1\}$ and $a_{i,\sigma(i)} \in \mathbb{Z}$. So $f(A)$ is a sum of integer terms. □

For example, with

$$A = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

we have

$$S_2 = \{(1,2),(2,1)\} \quad , \quad sign((1,2)) = 1 \quad , \quad sign((2,1)) = -1$$

and

$$f(A) = a_{1,1}a_{2,2} + (-1)a_{1,2}a_{21}$$

which is indeed the formula for the determinant of a 2×2 matrix.

11.12. Geometric properties for $\mathbb{F} = \mathbb{R}$

Suppose $\mathbb{F} = \mathbb{R}$. Consider the 2×2 matrix

$$A = \begin{bmatrix} 2 & 1 \\ 1 & 3 \end{bmatrix}$$

This matrix A is invertible and has

$$\det(A) = 5$$

Note that A maps the standard basis vectors $e_1 = (1,0)$ and $e_2 = (0,1)$ to the vectors $y_1 = (2,1)$ and $y_2 = (1,3)$, respectively. This is shown in Fig. 11.1 with e_1, y_1 in red and e_2, y_2 in blue. Moreover, A maps each point in the unit square $[0,1]^2$ to a unique point in the shaded parallelogram (Fig. 11.1). One can use basic geometry to show that the area of the parallelogram in Fig. 11.1 is 5, which is exactly $\det(A)$.

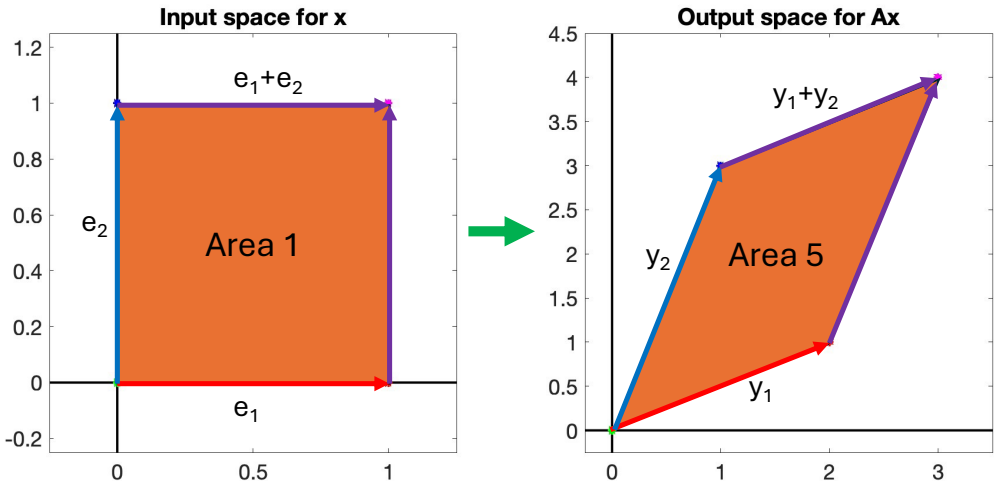


FIGURE 11.1. Mapping the unit square to a parallelogram of area 5 via matrix A (with $|\det(A)| = 5$). The edges satisfy $y_1 = Ae_1$ and $y_2 = Ae_2$.

Now consider the matrix B formed by swapping the columns of A:

$$B = \begin{bmatrix} 1 & 2 \\ 3 & 1 \end{bmatrix}$$

Swapping the columns multiplies the determinant by -1, and so $\det(B) = -5$. The corresponding map of the unit square through the matrix B results in a parallelogram. It is the same parallelogram as before, with the only difference that the orientation is flipped: The basis vector e_1 now maps to y_2; the basis vector e_2 now maps to y_1. The area of the parallelogram is still 5, which is the absolute value of $\det(B)$.

This turns out to be true in general: A 2×2 real matrix A maps the unit square $[0,1]^2$ to a parallelogram in $\mathbb{R}^2$ with area exactly equal to $|\det(A)|$. [Remarkably, if A has all integer entries, the area of the corresponding parallelogram is an integer, since $\det(A)$ is an integer.] The two edges of the parallelogram are Ae_1 and Ae_2, which are the same as the two columns of A. The 2-dimensional parallelogram degenerates into a 1-dimensional line segment (with zero area) if and only if one of the columns is a multiple of the other, which is precisely when the determinant is zero.

This has an important consequence: Take any 2-dimensional blob-like subset $S \subseteq \mathbb{R}^2$ that has a well defined and finite area (Fig. 11.2). Define AS as the set of points that each point in S maps to:

$$AS = \{Ax : x \in S\}$$

Then

$$area(AS) = area(S) \cdot |\det(A)|$$

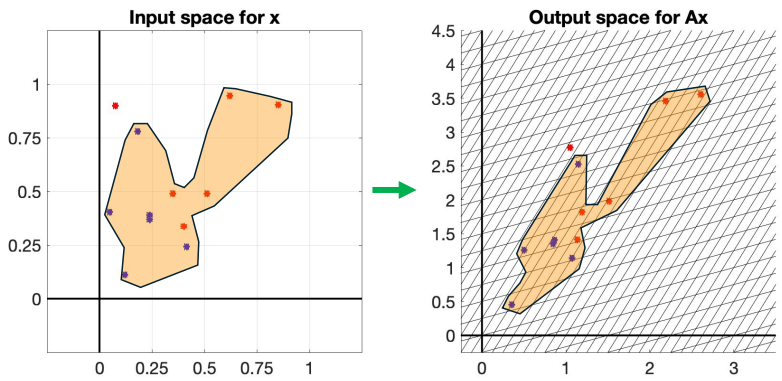


FIGURE 11.2. Mapping a Blob to a Blob with 5 times the area via matrix A with $|\det(A)| = 5$.

That is, the matrix A scales area by $|\det(A)|$. This is because $\mathbb{R}^2$ can be decomposed into tiny disjoint squares, each of which uniquely maps to a tiny parallelogram. Each tiny parallelogram has area equal to the area of the tiny square multiplied by $|\det(A)|$. Our subset S can be approximated by a union of the tiny squares, so its output AS can be approximated by the union of the tiny parallelograms (Fig. 11.2). The approximation becomes exact in the limit as the square size goes to zero. This argument holds for any subset S that can be approximated by tiny squares. Such subsets of $\mathbb{R}^2$ are *Borel measurable* subsets.

This property holds for $n \geq 2$: Let A be a real $n \times n$ matrix. Let $S \subseteq \mathbb{R}^n$ be a Borel measurable subset with finite multi-dimensional volume, denoted $Vol(S)$. Then AS is also Borel measurable and has volume scaled by $|\det(A)|$:

$$Vol(AS) = Vol(S) \cdot |\det(A)|$$

This volume-scaling property is used to calculate probability density functions in probability theory: Suppose A is a real $n \times n$ invertible matrix. Let $X = (X_1, \ldots, X_n)$ be a random vector with probability density function (PDF) $f_X(x)$ for $x \in \mathbb{R}^n$. Fix $b \in \mathbb{R}^n$ and define $Y = AX + b$. Then the PDF of Y is

$$f_Y(y) = \frac{f_X(A^{-1}(y-b))}{|\det(A)|} \quad \forall y \in \mathbb{R}^n$$

This is used to establish the *joint Gaussian PDF*: Let $X = (X_1, \ldots, X_n)$ be independent and identically distributed Gaussian random variables, each with mean 0 and variance 1. The joint PDF of $(X_1, \ldots, X_n)$ is just the product of the marginal (Gaussian) PDFs:

$$f_X(x) = \prod_{i=1}^{n} \frac{\exp(-\frac{1}{2}x_i^2)}{\sqrt{2\pi}} = \frac{\exp\left(-\frac{1}{2}x^\top x\right)}{(2\pi)^{n/2}} \quad \forall x \in \mathbb{R}^n$$

With A a real and invertible $n \times n$ matrix, we define $Y = AX + b$ and obtain for all $y \in \mathbb{R}^n$:

$$\begin{aligned} f_Y(y) &= \frac{f_X(A^{-1}(y-b))}{|\det(A)|} \\ &= \frac{\exp\left(-\frac{1}{2}(A^{-1}(y-b))^\top A^{-1}(y-b)\right)}{(2\pi)^{n/2}|\det(A)|} \\ &= \frac{\exp\left(-\frac{1}{2}(y-b)^\top (A^{-1})^\top A^{-1}(y-b)\right)}{(2\pi)^{n/2}|\det(A)|} \end{aligned} \tag{11.17}$$

This formula is often written in terms of the *covariance matrix* $\Lambda = AA^\top$, so

$$\begin{aligned} \det(\Lambda) &= \det(A)\det(A^\top) = \det(A)^2 \\ \Lambda^{-1} &= (AA^\top)^{-1} = (A^\top)^{-1}A^{-1} = (A^{-1})^\top A^{-1} \end{aligned}$$

Substituting these identities into (11.17) gives

$$\boxed{f_Y(y) = \frac{\exp\left(-\frac{1}{2}(y-b)^\top \Lambda^{-1}(y-b)\right)}{(2\pi)^{n/2}\det(\Lambda)^{1/2}} \quad \forall y \in \mathbb{R}^n}$$

which is the PDF for a jointly Gaussian random vector Y with mean $\mathbb{E}[Y] = b$ and covariance $\Lambda = \mathbb{E}\left[(Y-b)(Y-b)^\top\right]$, assuming Λ is invertible. The most natural way to simulate jointly Gaussian random variables $(Y_1, \ldots, Y_n)$ with a given mean and covariance b and Λ is to find any invertible matrix A that satisfies $\Lambda = AA^\top$, then define $Y = AX + b$ where $X = (X_1, \ldots, X_n)$ are n independent and identically distributed Gaussian random variables with mean 0 and variance 1. Note that every covariance matrix Λ is symmetric and positive semidefinite. It turns out that every real, symmetric, and positive semidefinite matrix Λ indeed has a matrix A such that $\Lambda = AA^\top$. The matrix A can be viewed as a type of "square root" for Λ. Such a square root matrix A can be found using the theory of eigenvalues and eigenvectors (treated in the next chapters). Specifically, we can *diagonalize* the (symmetric) matrix Λ by finding a basis of (orthonormal) eigenvectors, let S be the matrix with columns equal to these eigenvectors, then $\Lambda = S\Sigma S^\top$ where Σ is a diagonal matrix with all nonnegative entries. Then we can choose $A = S\Sigma^{1/2}$, where $\Sigma^{1/2}$ is the diagonal matrix formed by taking square roots of all diagonal entries of Σ. Alternatively, by observing $S^\top S = I$, we can use the symmetric square root $A = S\Sigma^{1/2}S^\top$ (so $A^2 = AA^\top = \Lambda$).

11.13. Determinant of linear operator

Let V be a vector space over a field $\mathbb{F}$. Assume $\dim(V) = n$ for some $n \in \mathbb{N}$. Let $T : V \to V$ be a linear operator. Let $\{v_1, \ldots, v_n\}$ and $\{q_1, \ldots, q_n\}$ be two different bases for V. Let M_1 be the matrix $Mat(T)$ under basis $\{v_1, \ldots, v_n\}$; let M_2 be the matrix $Mat(T)$ under basis

$\{q_1, \ldots, q_n\}$. It can be shown that $M_2 = PM_1P^{-1}$ for some invertible matrix P, and

$$\det(M_1) = \det(M_2)$$
$$\text{trace}(M_1) = \text{trace}(M_2)$$

Thus, the matrix of linear operator T has the same determinant and trace regardless of the basis used to form the matrix. We define the *determinant of linear operator T* and the *trace of linear operator T* as the determinant and trace of $Mat(T)$ under any basis for V.

11.14. Exercises

EXERCISE 11.1. *(Determinants of structured matrices) Using $\mathbb{F} = \mathbb{R}$, use the structure of A to compute* $\det(A)$ *for:*

a)

$$A = \begin{bmatrix} 3 & 8 & 1 & 1 \\ 0 & -1 & 4 & 80 \\ 0 & 0 & 2 & -100 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

b)

$$A = \begin{bmatrix} 4 & -2 \\ 1 & 2 \end{bmatrix}$$

c)

$$A = \begin{bmatrix} 3 & -2 & 4 & 7 & 9 \\ 8 & 8 & 8 & 7 & 8 \\ 0 & 0 & 2 & 4 & 0 \\ 0 & 0 & 0 & 7 & 1 \\ 0 & 0 & 0 & 0 & -2 \end{bmatrix}$$

d)

$$A = \begin{bmatrix} 5 & 1 & 0 & 1 & 1 & 7 & 9 \\ 9 & 1 & -1 & -45 & 3 & 0 & 2 \\ 0 & 0 & 3 & -2 & 73 & 0 & 0 \\ 0 & 0 & 0 & -10 & 9 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & -2 & 9 & 2 & 5 \\ 0 & 0 & 7 & -2 & 9 & -8 & 7 \end{bmatrix}$$

EXERCISE 11.2. *(Determinant by Gaussian elimination) For* $\mathbb{F} = \mathbb{R}$*, use Gaussian elimination to compute* $\det(A)$ *when*

$$A = \begin{bmatrix} 1 & 3 & -2 & 1 \\ 2 & 2 & 2 & 1 \\ 0 & 1 & 0 & 2 \\ 3 & -1 & 2 & 1 \end{bmatrix}$$

EXERCISE 11.3. *(Determinant by Gaussian elimination in binary field) Use Gaussian elimination in the binary field* $\mathbb{F}_2 = \{0, 1\}$ *(with mod 2 arithmetic) to compute* $\det(A) \in \{0, 1\}$ *for*

a)

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

b)

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

c)

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

EXERCISE 11.4. *(Determinant of integer matrix) Consider* $\mathbb{F} = \mathbb{R}$*. Show that if* A *is a square matrix and all entries are integers, then* $\det(A)$ *is an integer.*

EXERCISE 11.5. *Fix* p *as a prime. Suppose* A *is a* $n \times n$ *matrix with all entries in* $\{0, 1, \ldots, p-1\}$*. Let* $x = \det(A)$ *be taken when* $\mathbb{F} = \mathbb{R}$*. We know from the previous problem that* x *is an integer. Let* $y = \det(A)$ *when* $\mathbb{F} = \mathbb{F}_p = \{0, 1, \ldots, p-1\}$ *with mod* p *arithmetic. Argue that* $y = \mod(x, p)$*. Conclude that if* A *is invertible under field* $\mathbb{F}_p$*, it must also be invertible under field* $\mathbb{R}$*.*

EXERCISE 11.6. *Give a specific example of a* 3×3 *matrix* A *that has all entries either* 0 *or* 1*, such that* $\det(A) \neq 0$ *when* $\mathbb{F} = \mathbb{R}$*, but* $\det(A) = 0$ *when* $\mathbb{F} = \mathbb{F}_2 = \{0, 1\}$*.*

EXERCISE 11.7. *(Row 2 cofactor expansion for determinant of* 3×3*) Fix* $\mathbb{F}$ *as a field and let*

$$A = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

a) Repeat a derivation similar to (11.3)-(11.5)*, but for an expansion on row 2. Your first line should use linearity in row 2 to obtain*

$$\begin{aligned} \det(A) &= d \det\left(\begin{bmatrix} a & b & c \\ 1 & 0 & 0 \\ g & h & i \end{bmatrix}\right) + e \det\left(\begin{bmatrix} a & b & c \\ 0 & 1 & 0 \\ g & h & i \end{bmatrix}\right) + f \det\left(\begin{bmatrix} a & b & c \\ 0 & 0 & 1 \\ g & h & i \end{bmatrix}\right) \\ &= \textit{Expression 2} \\ &= \textit{Expression 3} \end{aligned}$$

Your Expression 2 should swap the first and second rows of each matrix; your Expression 3 should move each corresponding column to the very left.

b) Use your expression, as well as the resulting expression of (11.3)-(11.5)*, to show they give the same answer when* $\mathbb{F} = \mathbb{R}$ *and*

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 4 & 4 \\ -1 & 1 & 1 \end{bmatrix}$$

EXERCISE 11.8. *(Existence of a determinant) Fix* $\mathbb{F}$ *as a field. Fix* n *a positive integer (assume* $n \geq 2$*). Consider the function* $f : \mathbb{F}^{n \times n} \to \mathbb{F}$ *defined by*

$$f(A) = \sum_{\sigma \in S_n} sign(\sigma) \prod_{i=1}^{n} a_{i,\sigma(i)}$$

where the set S_n *is defined in Section 11.11. We want to show* f *satisfies the three axioms of determinants (and so* $f(A) = \det(A)$*).*

a) Show Axiom 1 holds: $f(I) = 1$*. Hint: If* $\sigma \neq (1, 2, 3, \ldots, n)$ *then* $\prod_{i=1}^{n} a_{\sigma(i),i} = 0$*.*

b) Show Axiom 3 holds for row 1 (the same argument can be shown for any row i, but this exercise considers only row 1 for simplicity).

c) Define $g : \mathbb{F}^{n \times n} \to \mathbb{F}$ *by*

$$g(A) = \sum_{\sigma \in S_n} sign(\sigma) \prod_{i=1}^{n} a_{\sigma(i),i}$$

Explain why $g(A) = f(A)$ *for all* $n \times n$ *matrices* A*. Hint: Observe that for each permutation*

$\sigma \in S_n$ *we have* $\prod_{i=1}^{n} a_{i,\sigma(i)} = \prod_{j=1}^{n} a_{\sigma^{-1}(j),j}$ *where* $\sigma^{-1} \in S_n$ *is the inverse permutation, meaning that* $\sigma^{-1}(j)$ *is the unique index* $i \in \{1, \ldots, n\}$ *for which* $\sigma(i) = j$.

d) Show that if A *has a repeated row, then* $g(A) = 0$, *hence* $f(A) = 0$. *Hint: Fix distinct rows* $r, s \in \{1, \ldots, n\}$. *Observe that* $S_n = Y_n \cup Z_n$ *where* Y_n *is the set of permutations* $\sigma \in S_n$ *for which* r *is listed before* s, Z_n *is the set of permutations for which* s *is listed before* r. *Each permutation* $\sigma \in Y_n$ *corresponds to a permutation* $\hat{\sigma} \in Z_n$ *that preserves all orderings of* σ, *except for the ordering of* r *and* s, *and of course* $sign(\hat{\sigma}) = -sign(\sigma)$.

EXERCISE 11.9. *(Quadratic equations with finite fields) Let* $\mathbb{F}$ *be a finite field (so it has a finite number of scalars). This exercise shows there are quadratic equations of the form* $x^2 + bx + c = 0$ *that have no solution. We separately treat the cases when the field satisfies* $1 \neq -1$, *and when it satisfies* $1 = -1$.

a) Show that if a function $f : \mathbb{F} \to \mathbb{F}$ *is surjective (meaning for all* $s \in \mathbb{F}$ *there is an* $x \in \mathbb{F}$ *such that* $f(x) = s$), *then it is also injective (meaning it is one-to-one, that is, if scalars* x, y *satisfy* $f(x) = f(y)$ *then we must have* $x = y$).

b) Suppose the field satisfies $1 \neq -1$. *Consider the function* $f : \mathbb{F} \to \mathbb{F}$ *defined by* $f(x) = x^2$ *for all* $x \in \mathbb{F}$. *Show that* f *is not surjective. Use this to create a quadratic equation of the form* $x^2 + bx + c = 0$ *that has no root.*

c) Now suppose the field satisfies $1 = -1$. *Define* $g : \mathbb{F} \to \mathbb{F}$ *by* $g(x) = x^2 + x$ *for all* $x \in \mathbb{F}$. *Show* g *is not surjective. Use this to find a quadratic equation of the form* $x^2 + bx + c = 0$ *with no root.*

d) Give one example of a finite field that satisfies $1 = -1$ *(there are many such fields).*

CHAPTER 12.

Eigenvalues

Let $\mathbb{F}$ be a field. Let A be a square matrix with entries in $\mathbb{F}$. We want to know if there is a scalar $\lambda \in \mathbb{F}$ and a nonzero vector $v \in \mathbb{F}^n$ such that

$$Av = \lambda v$$

This question is only interesting if $v \neq 0$, since the equation is trivially true if $v = 0$. On the other hand, the question is still interesting when $\lambda = 0$ (in that case the question reduces to asking whether or not A is invertible). Observe that if I is the identity matrix of the same size as A then

$$(Av = \lambda v) \iff (A - \lambda I)\, v = 0$$

DEFINITION 12.1. *A scalar $\lambda \in \mathbb{F}$ is an* eigenvalue *of a $n \times n$ matrix A if there is a nonzero vector $v \in \mathbb{F}^n$ such that*

$$Av = \lambda v$$

The nonzero vector v is called an eigenvector of A. *(Note that an eigenvector v must be nonzero, but an eigenvalue λ is allowed to be zero.) The pair (λ, v) is called an* eigenpair of A.

For example, consider

$$A = \begin{bmatrix} 2 & 2 \\ 0 & 7 \end{bmatrix} \quad , \quad v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 2 \\ 5 \end{bmatrix} \quad , \quad \tilde{v}_2 = \begin{bmatrix} 20 \\ 50 \end{bmatrix}$$

Then

$$\begin{bmatrix} 2 & 2 \\ 0 & 7 \end{bmatrix}\begin{bmatrix} 1 \\ 0 \end{bmatrix} = 2\begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad \text{[so 2 is an eigenvalue of } A \text{ with eigenvector } v_1\text{]}$$

$$\begin{bmatrix} 2 & 2 \\ 0 & 7 \end{bmatrix}\begin{bmatrix} 2 \\ 5 \end{bmatrix} = 7\begin{bmatrix} 2 \\ 5 \end{bmatrix} \quad \text{[so 7 is an eigenvalue of } A \text{ with eigenvector } v_2\text{]}$$

So $(2, v_1)$ and $(7, v_2)$ are eigenpairs of A. Notice that $(7, \tilde{v}_2)$ is another eigenpair of A. In general, multiplying an eigenvector by a nonzero scalar creates another eigenvector.

12.0.1. Repeated multiplication by A

Observe that if λ is an eigenvalue of A with eigenvector v then

$$Av = \lambda v \tag{12.1}$$

Multiplying both sides of (12.1) on the left by A gives

$$\begin{aligned} A^2 v &= A(\lambda v) \\ &= \lambda A v \\ &= \lambda^2 v \end{aligned}$$

where the final equality substitutes (12.1). Thus

$$A^2 v = \lambda^2 v$$

Multiplying both sides of this on the left by A gives

$$\begin{aligned} A^3 v &= A(\lambda^2 v) \\ &= \lambda^2 A v \\ &= \lambda^3 v \end{aligned}$$

where the final equality substitutes (12.1). In this way, we have

$$A^k v = \lambda^k v \quad \forall k \in \{1, 2, 3, \ldots\} \tag{12.2}$$

If $k = 0$ we define $A^0 = I$ and $\lambda^0 = 1$, which makes formula (12.2) work for $k = 0$. [If A is invertible we can define A^{-k} as $(A^{-1})^k$ for $k \in \mathbb{N}$. With this definition, the reader is invited to show that if A is invertible then formula (12.2) extends to all integers k.]

12.0.2. Basic consequences

LEMMA 12.1. *Let (λ, v) be an eigenpair of a matrix A. Then for each $\theta \in \mathbb{F}$ such that $\theta \neq 0$, $(\lambda, \theta v)$ is also an eigenpair of A.*

PROOF. If (λ, v) is an eigenpair then $v \neq 0$ and $Av = \lambda v$. Multiplying this by θ gives $A(\theta v) = \lambda(\theta v)$, where $\theta v \neq 0$ since $\theta \neq 0$. □

LEMMA 12.2. *Let A be a square matrix A. Then*

$$(A \text{ is invertible}) \iff (Null(A) = \{0\}) \iff (0 \text{ is not an eigenvalue of } A)$$

$$(A \text{ is not invertible}) \iff (Null(A) \neq \{0\}) \iff (0 \text{ is an eigenvalue of } A)$$

PROOF. See Exercise 12.2. □

12.1. Necessary and sufficient condition

LEMMA 12.3. *Let A be a $n \times n$ matrix with entries in some field $\mathbb{F}$.*

a) A scalar $\lambda \in \mathbb{F}$ is an eigenvalue of A if and only if $\det(A - \lambda I) = 0$.

b) Suppose λ is an eigenvalue of A. Then v is an eigenvector of A (with eigenvalue λ) if and only if $v \neq 0$ and $v \in Null(A - \lambda I)$.

PROOF. To prove (a), suppose λ is an eigenvalue of A. Then there is a nonzero vector v such that $Av = \lambda v$. Then $(A - \lambda I)v = 0$ and so

$$v \in Null(A - \lambda I)$$

which means $A - \lambda I$ is a square matrix with a nontrivial null space. So $A - \lambda I$ is not invertible and $\det(A - \lambda I) = 0$. Conversely, suppose λ satisfies $\det(A - \lambda I) = 0$. Then the square matrix $A - \lambda I$ is not invertible. So it has a nonzero vector v in its null space, meaning $(A - \lambda I)v = 0$. So $Av = \lambda v$ and λ is an eigenvalue. This proves part (a).

To prove (b), suppose λ is an eigenvalue of A. If v is a corresponding eigenvector then $v \neq 0$ and $Av = \lambda v$. So $(A - \lambda I)v = 0$, meaning v is a nonzero vector in $Null(A - \lambda I)$. Conversely, if v is a nonzero vector in $Null(A - \lambda I)$, then $(A - \lambda I)v = 0$ and so $Av = \lambda v$, meaning v is an eigenvector of A corresponding to eigenvalue λ. □

LEMMA 12.4. *Let A be a square matrix. Let λ be a scalar. Then*

$$(\lambda \text{ is an eigenvalue of } A) \iff (\lambda \text{ is an eigenvalue of } A^\top)$$

PROOF. This follows directly from the previous lemma (see Exercise 12.12). □

12.1.1. Examples

1. **Problem:** Consider field $\mathbb{R}$ and

$$A = \begin{bmatrix} 2 & 2 & 0 \\ 0 & -7 & 0 \\ 0 & 1 & 3 \end{bmatrix} \quad , \quad \lambda = -7 \quad , \quad v = \begin{bmatrix} ? \\ ? \\ ? \end{bmatrix}$$

It can be shown that -7 is an eigenvalue of A. Find an eigenvector v corresponding to eigenvalue -7.

Solution: Using $\lambda = -7$, we have

$$A - (-7)I = \begin{bmatrix} 9 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 10 \end{bmatrix}$$

Using Gaussian elimination to find a vector $v = (x, y, z)$ to solve $(A - (-7)I)v = 0$ gives

$$\begin{bmatrix} 9 & 2 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 10 \end{bmatrix} \overset{(a)}{\rightarrow} \begin{bmatrix} 1 & 2/9 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 10 \end{bmatrix} \overset{(b)}{\rightarrow} \begin{bmatrix} 1 & 2/9 & 0 \\ 0 & 1 & 10 \\ 0 & 0 & 0 \end{bmatrix} \overset{(c)}{\rightarrow} \begin{bmatrix} \boxed{1} & 0 & -20/9 \\ 0 & \boxed{1} & 10 \\ 0 & 0 & 0 \end{bmatrix}$$

where step (a) multiplies row 1 by 1/9; step (b) swaps rows 2 and 3; step (c) multiplies row 2 by $-2/9$ and adds into row 1. There is one free variable (variable z). All solutions $v = (x, y, z)$ satisfy

$$x + (-20/9)z = 0$$
$$y + 10z = 0$$

and so

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} (20/9)z \\ -10z \\ z \end{bmatrix} = z \begin{bmatrix} (20/9) \\ -10 \\ 1 \end{bmatrix}$$

so

$$Null(A - (-7)I) = Span\{(20/9, -10, 1)\}$$

It follows that there are *infinitely many* valid eigenvectors v we can choose. All of them

are nonzero scalar multiples of $(20/9, -10, 1)$. We only need one. Lets choose

$$\boxed{v = (20/9, -10, 1)}$$

(Alternatively, if we wanted an all-integer eigenvector, we could choose $(20, -90, 9)$.) We can test this by showing it satisfies $Av = (-7)v$:

$$\underbrace{\begin{bmatrix} 2 & 2 & 0 \\ 0 & -7 & 0 \\ 0 & 1 & 3 \end{bmatrix}}_{A}\underbrace{\begin{bmatrix} 20/9 \\ -10 \\ 1 \end{bmatrix}}_{v} = \begin{bmatrix} -140/9 \\ 70 \\ -7 \end{bmatrix} = (-7)\underbrace{\begin{bmatrix} 20/9 \\ -10 \\ 1 \end{bmatrix}}_{v}$$

2. **Problem:** Consider field $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic and

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

Find all eigenvalues of A. For each eigenvalue, find a corresponding eigenvector.

Solution: There are only two scalars in this field, and hence only two candidates for eigenvalues. To find all eigenvalues $\lambda \in \{0, 1\}$, we consider

$$\det(A - \lambda I) = \det\left(\begin{bmatrix} 1-\lambda & 1 \\ 1 & 1-\lambda \end{bmatrix}\right) = (1-\lambda)^2 - 1 = -2\lambda + \lambda^2 = \lambda^2 = \lambda$$

where we have used $2 = 0$ and $\lambda^2 = \lambda$ in the binary field. So

$$(\det(A - \lambda I) = 0) \iff (\lambda = 0)$$

So $\lambda = 0$ is the only eigenvalue. Using $\lambda = 0$ we have

$$A - (0)I = A = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$$

Using Gaussian elimination to find $Null(A)$ gives

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

It is clear that

$$Null(A) = \{(x, y) \in \mathbb{F}_2^2 : x + y = 0\} = \{(x, y) \in \mathbb{F}_2^2 : x = y\} = Span\{(1, 1)\}$$

The only possible eigenvector is $(1, 1)$. That is

$$\boxed{\lambda = 0, v = (1, 1)}$$

It is easy to verify that $Av = 0v$.

3. **Problem:** Consider field $\mathbb{R}$. Find the polynomial $\det(A - \lambda I)$ and find all eigenvalues of A for

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & 0 \\ 3 & 0 & 5 \end{bmatrix}$$

Solution: We have

$$A - \lambda I = \begin{bmatrix} 1 - \lambda & 1 & 1 \\ 2 & 1 - \lambda & 0 \\ 3 & 0 & 5 - \lambda \end{bmatrix}$$

Using the cofactor formula for determinants by expanding on the second row gives

$$\begin{aligned} \det(A - \lambda I) &= 2(-1)[(5 - \lambda) - 0] + (1 - \lambda)(1)[(1 - \lambda)(5 - \lambda) - 3] + 0 \\ &= -\lambda^3 + 7\lambda^2 - 6\lambda - 8 \end{aligned}$$

So we obtain a degree 3 polynomial:

$$\boxed{\det(A - \lambda I) = -\lambda^3 + 7\lambda^2 - 6\lambda - 8}$$

To find the roots $\lambda_1, \lambda_2, \lambda_3$, we can either factor by hand (if we notice that 2 is a root) or we can use a computer:

$$\det(A - \lambda I) = -(\lambda - 2)\left(\lambda - \frac{5 + \sqrt{41}}{2}\right)\left(\lambda - \frac{5 - \sqrt{41}}{2}\right)$$

$$\begin{aligned} \lambda_1 &= 2 \\ \lambda_2 &= \frac{5 + \sqrt{41}}{2} = 5.701562118716424 \\ \lambda_3 &= \frac{5 - \sqrt{41}}{2} = -0.701562118716424 \end{aligned}$$

The displayed numerical values are finite precision representations of a real number. We can verify that we have made no mistakes by using a computer. In MATLAB:

```
» eig(A)
ans =
5.701562118716424 -0.701562118716425 2.000000000000000
```

4. **Problem:** With field $\mathbb{R}$, find all eigenvalues of the following matrix. For each eigenvalue, find a corresponding eigenvector.

$$A = \begin{bmatrix} 2 & 2 & 3 & 5 & 8 \\ 0 & -3 & 1 & 1 & 2 \\ 0 & 0 & 5 & 1 & 1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 4 & 4 \end{bmatrix}$$

Solution: We have

$$A - \lambda I = \left[\begin{array}{ccc|cc} 2-\lambda & 2 & 3 & 5 & 8 \\ 0 & -3-\lambda & 1 & 1 & 2 \\ 0 & 0 & 5-\lambda & 1 & 1 \\ \hline 0 & 0 & 0 & 2-\lambda & 2 \\ 0 & 0 & 0 & 4 & 4-\lambda \end{array}\right] \tag{12.3}$$

This has a block upper triangular structure so

$$\begin{aligned} \det(A - \lambda I) &= \det\left(\begin{bmatrix} 2-\lambda & 2 & 3 \\ 0 & -3-\lambda & 1 \\ 0 & 0 & 5-\lambda \end{bmatrix}\right) \det\left(\begin{bmatrix} 2-\lambda & 2 \\ 4 & 4-\lambda \end{bmatrix}\right) \\ &= (2-\lambda)(-3-\lambda)(5-\lambda)\left[(2-\lambda)(4-\lambda) - 8\right] \\ &= (2-\lambda)(-3-\lambda)(5-\lambda)(\lambda - 0)(\lambda - 6) \end{aligned}$$

so there are exactly five eigenvalues:

$$\lambda_1 = 2, \lambda_2 = -3, \lambda_3 = 5, \lambda_4 = 0, \lambda_5 = 6$$

For the first three eigenvalues, the block matrix structure suggests we can search for eigenvectors of the form $(x, y, z, 0, 0)$. On the other hand, eigenvectors for λ_4 and λ_5 will *not necessarily* have the form $(0, 0, 0, x, y)$. In other words, we use a simplifying

technique that holds only for the upper left block of our block triangular structure. For $\lambda_1, \lambda_2, \lambda_3$ we search for nonzero $(x, y, z) \in \mathbb{R}^3$ in the null space of

$$\begin{bmatrix} 2-\lambda_i & 2 & 3 \\ 0 & -3-\lambda_i & 1 \\ 0 & 0 & 5-\lambda_i \end{bmatrix} \tag{12.4}$$

- $\lambda_1 = 2$. It is easy to see that $v_1 = (1, 0, 0, 0, 0)$ is an eigenvector for $\lambda_1 = 2$.
- $\lambda_2 = -3$. We search for nonzero $(x, y, z) \in \mathbb{R}^3$ in the null space of the matrix in (12.4) given by substituting $\lambda_i = -3$. By Gaussian elimination

$$\begin{bmatrix} 5 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 8 \end{bmatrix} \to \begin{bmatrix} 5 & 2 & 3 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

  So we need $(x, y, z) \neq (0, 0, 0)$ to satisfy

$$\begin{aligned} 5x + 2y + 3z &= 0 \\ z &= 0 \end{aligned}$$

  There is one free variable (y). To ensure our solution is not all-zero, we can set $y = 1$ to obtain $(x, y, z) = (-2/5, 1, 0)$. If we want, we can scale this by 5 to get an all-integer solution (which is not required, but may look nicer). So we use $(x, y, z) = (-2, 5, 0)$ and so our eigenvector for λ_2 is

$$v_2 = (-2, 5, 0, 0, 0)$$

- $\lambda_3 = 5$. Substituting $\lambda_i = 5$ into (12.4) gives

$$\begin{bmatrix} -3 & 2 & 3 \\ 0 & -8 & 1 \\ 0 & 0 & 0 \end{bmatrix}$$

  We search for nonzero $(x, y, z) \in \mathbb{R}^3$ in the null space of this matrix. We need

$$\begin{aligned} -3x + 2y + 3z &= 0 \\ -8y + z &= 0 \end{aligned}$$

  We can choose $z = 1$ and then $(x, y, z) = (13/12, 1/8, 1)$. If we want, we can scale this

by 24 to obtain the integer solution (which looks nicer): $(x, y, z) = (26, 3, 24)$. So our eigenvector for λ_3 is

$$v_3 = (26, 3, 24, 0, 0)$$

- $\lambda_4 = 0$. We need any nonzero vector $v_4 = (x_1, x_2, x_3, x_4, x_5) \in Null(A)$. By Gaussian elimination

$$\begin{bmatrix} 2 & 2 & 3 & 5 & 8 \\ 0 & -3 & 1 & 1 & 2 \\ 0 & 0 & 5 & 1 & 1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 4 & 4 \end{bmatrix} \rightarrow \begin{bmatrix} 2 & 2 & 3 & 5 & 8 \\ 0 & -3 & 1 & 1 & 2 \\ 0 & 0 & 5 & 1 & 1 \\ 0 & 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

We have one free variable x_5. Choose $x_5 = 1$. Then

$$\begin{aligned} & x_5 = 1 \\ & 2x_4 + 2x_5 = 0 \implies x_4 = -1 \\ & 5x_3 + x_4 + x_5 = 0 \implies x_3 = 0 \\ & -3x_2 + x_3 + x_4 + 2x_5 = 0 \implies x_2 = 1/3 \\ & 2x_1 + 2x_2 + 3x_3 + 5x_4 + 8x_5 = 0 \implies x_1 = -11/6 \end{aligned}$$

We can scale by 6 to obtain $v_4 = (-11, 2, 0, -6, 6)$.

- $\lambda_5 = 6$. Substituting $\lambda = 6$ into (12.3) gives

$$A - 6I = \begin{bmatrix} -4 & 2 & 3 & 5 & 8 \\ 0 & -9 & 1 & 1 & 2 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & -4 & 2 \\ 0 & 0 & 0 & 4 & -2 \end{bmatrix}$$

We want a nonzero vector $v_5 = (x_1, x_2, x_3, x_4, x_5)$ in the null space:

$$\begin{bmatrix} -4 & 2 & 3 & 5 & 8 \\ 0 & -9 & 1 & 1 & 2 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & -4 & 2 \\ 0 & 0 & 0 & 4 & -2 \end{bmatrix} \rightarrow \begin{bmatrix} -4 & 2 & 3 & 5 & 8 \\ 0 & -9 & 1 & 1 & 2 \\ 0 & 0 & -1 & 1 & 1 \\ 0 & 0 & 0 & -4 & 2 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

Let the free variable be $x_5 = 1$. Then

$$\begin{aligned}
-4x_4 + 2x_5 = 0 &\implies x_4 = 1/2 \\
-x_3 + x_4 + x_5 = 0 &\implies x_3 = 3/2 \\
-9x_2 + x_3 + x_4 + 2x_5 = 0 &\implies x_2 = 4/9 \\
-4x_1 + 2x_2 + 3x_3 + 5x_4 + 8x_5 = 0 &\implies x_1 = 143/36
\end{aligned}$$

Scaling by 36 (only for convenience) gives $v_5 = (143, 16, 54, 18, 36)$.

It is standard to organize this information as follows: Collect all eigenvalues into a diagonal matrix Λ, and the corresponding eigenvectors as columns in a matrix S:

$$\Lambda = \begin{bmatrix} 2 & 0 & 0 & 0 & 0 \\ 0 & -3 & 0 & 0 & 0 \\ 0 & 0 & 5 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 6 \end{bmatrix}, S = \begin{bmatrix} | & | & | & | & | \\ v_1 & v_2 & v_3 & v_4 & v_5 \\ | & | & | & | & | \end{bmatrix} = \begin{bmatrix} 1 & -2 & 26 & -11 & 143 \\ 0 & 5 & 3 & 2 & 16 \\ 0 & 0 & 24 & 0 & 54 \\ 0 & 0 & 0 & -6 & 18 \\ 0 & 0 & 0 & 6 & 36 \end{bmatrix}$$

We should now test that our computations are correct:

$$Av_i = \lambda_i v_i \quad \text{[correct for all } i \in \{1, 2, 3, 4, 5\}\text{]}$$

An equivalent and clever way to check our work is to verify

$$AS = S\Lambda \quad \text{[correct]} \tag{12.5}$$

In this case we have 5 distinct eigenvalues for our 5×5 matrix A. It turns out that the corresponding eigenvectors $\{v_1, v_2, v_3, v_4, v_5\}$ are linearly independent (see next section). So the matrix S must be invertible. Multiplying (12.5) on the right by S^{-1} gives $A = S\Lambda S^{-1}$, which can be verified by computer.

12.2. Linear independence property

THEOREM 12.1. *Let A be a square matrix with entries in some field $\mathbb{F}$.*

a) Suppose $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of A. Let $v_1, \ldots, v_k$ be corresponding eigenvectors (so $Av_i = \lambda_i v_i$ for all $i \in \{1, \ldots, k\}$). Then $\{v_1, \ldots, v_k\}$ are linearly independent.

b) If A has size $n \times n$, then it can have at most n distinct eigenvalues.

PROOF. To prove (a), we proceed by induction. It trivially holds for $k = 1$ because eigenvectors are nonzero. Suppose it holds for some positive integer $k \geq 1$. We show

it holds for $k+1$. Let $\lambda_1, \ldots, \lambda_{k+1}$ be distinct eigenvalues of A. Let $\{v_1, \ldots, v_{k+1}\}$ be the corresponding eigenvectors, so $Av_i = \lambda_i v_i$ for $i \in \{1, \ldots, k+1\}$. Suppose that

$$\sum_{i=1}^{k+1} a_i v_i = 0 \tag{12.6}$$

for some scalars $a_1, \ldots, a_{k+1}$. We want to show $a_i = 0$ for all $i \in \{1, \ldots, k+1\}$. From (12.6) we have

$$\sum_{i=1}^{k} a_i v_i = -a_{k+1} v_{k+1} \tag{12.7}$$

Multiplying both sides of (12.7) by A gives

$$\sum_{i=1}^{k} a_i \lambda_i v_i = -a_{k+1} \lambda_{k+1} v_{k+1}$$

On the other hand, multiplying both sides of (12.7) by λ_{k+1} gives

$$\sum_{i=1}^{k} a_i \lambda_{k+1} v_i = -a_{k+1} \lambda_{k+1} v_{k+1}$$

Equating the above two equalities gives

$$\sum_{i=1}^{k} a_i (\lambda_i - \lambda_{k+1}) v_i = 0$$

Since $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues, we know by the induction assumption that $\{v_1, \ldots, v_k\}$ are linearly independent, and so the above equation immediately implies

$$a_i (\lambda_i - \lambda_{k+1}) = 0 \quad \forall i \in \{1, \ldots, k\}$$

Since $\lambda_i \neq \lambda_{k+1}$ for $i \in \{1, \ldots, k\}$, we know $a_i = 0$ for all $i \in \{1, \ldots, k\}$. Thus, (12.6) becomes

$$a_{k+1} v_{k+1} = 0$$

and since $v_{k+1} \neq 0$, we must have $a_{k+1} = 0$, completing the proof of part (a).

Part (b) follows immediately from the fact that we cannot have more than n linearly independent vectors in $\mathbb{F}^n$. □

12.3. Eigenvalues of upper triangular matrices

Suppose U is a $n \times n$ upper-triangular matrix with entries in $\mathbb{F}$. Let $d_1, \ldots, d_n$ be the diagonal entries:

$$U = \begin{bmatrix} d_1 & * & * & \cdots & * \\ 0 & d_2 & * & \cdots & * \\ 0 & 0 & d_3 & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & * \\ 0 & 0 & \cdots & \cdots & d_n \end{bmatrix}$$

For each $\lambda \in \mathbb{F}$ we have

$$\det(U - \lambda I) = \det\left(\begin{bmatrix} d_1 - \lambda & * & * & \cdots & * \\ 0 & d_2 - \lambda & * & \cdots & * \\ 0 & 0 & d_3 - \lambda & \cdots & * \\ \vdots & \vdots & \vdots & \ddots & * \\ 0 & 0 & \cdots & \cdots & d_n - \lambda \end{bmatrix}\right) = (d_1 - \lambda)(d_2 - \lambda)\cdots(d_n - \lambda)$$

which uses the fact that the determinant of an upper triangular matrix is the product of its diagonal entries. It follows that a scalar λ is an eigenvalue of U if and only if $\prod_{i=1}^{n}(d_i - \lambda) = 0$. That is, λ is an eigenvalue of U if and only if $\lambda \in \{d_1, d_2, \ldots, d_n\}$. So all diagonal entries of U are eigenvalues, and U has no other eigenvalues. The same holds for lower triangular matrices. In summary:

- If U is a square upper triangular matrix, then its eigenvalues are its diagonal entries.
- If L is a square lower triangular matrix, then its eigenvalues are its diagonal entries.

For example, the $n \times n$ identity matrix I is upper triangular and has only 1s on its diagonal, and so its only eigenvalue is 1.

12.4. Computing eigenvectors

Finding the eigenvalues of a $n \times n$ matrix A involves solving $\det(A - \lambda I) = 0$. It turns out that this is always a degree n polynomial equation in λ. Once eigenvalues are known, it is easy to find the corresponding eigenvectors by Gaussian elimination.

1. Suppose $\mathbb{F} = \mathbb{R}$ and

$$A = \begin{bmatrix} 1 & 2 & 2 \\ 0 & 0 & 4 \\ 0 & 0 & -8 \end{bmatrix}$$

Since A is upper-triangular, we know the eigenvalues are $\lambda_1 = 1, \lambda_2 = 0, \lambda_3 = -8$. These are distinct and hence their corresponding eigenvectors must be linearly independent. Since the eigenvalue 1 appears in the upper left corner of the upper triangular matrix A, is easy to see that a corresponding eigenvector is

$$\lambda_1 = 1, \quad v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

and one can verify that $Av_1 = v_1$. The other eigenvectors are not as easy to find. Consider $\lambda_2 = 0$. To find $v_2 \neq 0$ that satisfies $Av_2 = 0 \cdot v_2$, we find $v_2 = (a, b, c) \neq 0$ to solve

$$\begin{bmatrix} 1 & 2 & 2 \\ 0 & 0 & 4 \\ 0 & 0 & -8 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

It is easy to see that a nonzero solution is $(a, b, c) = (2, -1, 0)$ and so we have

$$\lambda_2 = 0, \quad v_2 = \begin{bmatrix} 2 \\ -1 \\ 0 \end{bmatrix}$$

One can verify that $Av_2 = 0 \cdot v_2$. Finally, consider $\lambda_3 = -8$. We must find v_3 to solve $(A + 8I)v_3 = 0$, so we find $(a, b, c) \neq 0$ to solve

$$\begin{bmatrix} 9 & 2 & 2 \\ 0 & 8 & 4 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

There is one free variable c. To ensure the solution is nonzero, choose $c = 1$:

$$8b + 4 = 0 \implies b = -1/2$$

$$9a + 2b + 2 = 0 \implies a = \frac{-2 - 2(-1/2)}{9} = -1/9$$

Thus

$$\lambda_3 = -8, \quad v_3 = \begin{bmatrix} -1/9 \\ -1/2 \\ 1 \end{bmatrix}$$

and one can verify $Av_3 = (-8)v_3$. It is useful to collect this information in two square matrices Λ and S

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & -8 \end{bmatrix}$$

$$S = \begin{bmatrix} | & | & | \\ v_1 & v_2 & v_3 \\ | & | & | \end{bmatrix} = \begin{bmatrix} 1 & 2 & -1/9 \\ 0 & -1 & -1/2 \\ 0 & 0 & 1 \end{bmatrix}$$

Since $Av_i = \lambda_i v_i$ for $i \in \{1, 2, 3\}$, the rules of matrix multiplication give $AS = S\Lambda$, and since S is invertible we obtain $A = S\Lambda S^{-1}$ which, as shown in Section 12.7, is a *diagonalizable* form for the matrix A that holds whenever A has a basis of eigenvectors.

2. Consider $\mathbb{F} = \mathbb{R}$ and

$$A = \begin{bmatrix} 9 & 1 \\ -2 & 5 \end{bmatrix}$$

We first find the eigenvalues by solving $\det(A - \lambda I) = 0$, which means

$$(9 - \lambda)(5 - \lambda) - (1)(-2) = 0$$

and so $\lambda^2 - 14\lambda + 47 = 0$, which has exactly two roots:

$$\lambda_1 = 7 + \sqrt{2}$$
$$\lambda_2 = 7 - \sqrt{2}$$

It remains to find the corresponding eigenvectors v_1, v_2. We can save time by solving for them both at the same time. Let λ be an eigenvalue (either $\lambda = \lambda_1$ or $\lambda = \lambda_2$). We want to find a vector v in the null space of $A - \lambda I$, where

$$A - \lambda I = \begin{bmatrix} 9 - \lambda & 1 \\ -2 & 5 - \lambda \end{bmatrix}$$

Since we already know the null space is nontrivial (since λ is an eigenvalue), the second

equation is linearly dependent. Equivalently, since we know $9 - \lambda \neq 0$, we can perform Gaussian elimination:

$$\begin{bmatrix} 9-\lambda & 1 \\ -2 & 5-\lambda \end{bmatrix} \to \begin{bmatrix} 9-\lambda & 1 \\ 0 & 0 \end{bmatrix}$$

It suffices to find any nonzero vector (a, b) that satisfies

$$(9 - \lambda)a + b = 0$$

Letting $a = 1$ and then $b = \lambda - 9$, so we have

$$\lambda \implies v = \begin{bmatrix} 1 \\ \lambda - 9 \end{bmatrix}$$

So we have

$$\lambda_1 = 7 + \sqrt{2} \implies v_1 = \begin{bmatrix} 1 \\ \lambda_1 - 9 \end{bmatrix}$$

$$\lambda_2 = 7 - \sqrt{2} \implies v_2 = \begin{bmatrix} 1 \\ \lambda_2 - 9 \end{bmatrix}$$

and one can verify $Av_i = \lambda_i v_i$ for $i \in \{1, 2\}$. As before, we can collect the eigenpair information into matrices Λ and S:

$$\Lambda = \begin{bmatrix} 7+\sqrt{2} & 0 \\ 0 & 7-\sqrt{2} \end{bmatrix} \quad , \quad S = \begin{bmatrix} 1 & 1 \\ -2+\sqrt{2} & -2-\sqrt{2} \end{bmatrix}$$

and we indeed have the diagonalizable form $A = S\Lambda S^{-1}$.

12.5. Spectrum of a matrix

The *spectrum of a square matrix A*, written $\Lambda(A)$, is defined as the set of eigenvalues of A (this set is possibly empty). For example, consider the following matrices with $\mathbb{F} = \mathbb{R}$:

$$A = \begin{bmatrix} 2 & 9 \\ 0 & -4 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 7 & 9 & 0 \\ 0 & 7 & 1 \\ 0 & 0 & 5 \end{bmatrix} \quad , \quad C = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Then

$$\Lambda(A) = \{2, -4\}$$
$$\Lambda(B) = \{7, 5\}$$
$$\Lambda(C) = \phi \quad \text{(the empty set)}$$

The matrix C rotates all vectors in $\mathbb{R}^2$ by 90 degrees, so there are no eigenvectors in the real field. The equation $\det(C - \lambda I) = 0$ is $\lambda^2 + 1 = 0$, which indeed has no real solutions. However, the equation $\lambda^2 + 1 = 0$ has two complex solutions $\lambda_1 = i$ and $\lambda_2 = -i$. Therefore, if we use the complex field $\mathbb{F} = \mathbb{C}$ we get $\Lambda(C) = \{i, -i\}$.

12.6. Eigenvalues of a 2×2 matrix

Fix $\mathbb{F}$ as a field and consider the general 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

We know $\lambda \in \mathbb{F}$ is an eigenvalue of A if and only if $\det(A - \lambda I) = 0$, which holds if and only if $(a - \lambda)(d - \lambda) - bc = 0$, which holds if and only if

$$\lambda^2 - (a + d)\lambda + ad - bc = 0$$

For simplicity, suppose our field $\mathbb{F}$ satisfies $1 + 1 \neq 0$. Define $2 = 1 + 1$. Since we assume $2 \neq 0$, we know 2^{-1} exists and the quadratic formula gives all roots of this equation (if they exist):

$$\lambda_1 = \frac{(a + d) + \sqrt{(a - d)^2 + 4bc}}{2}$$
$$\lambda_2 = \frac{(a + d) - \sqrt{(a - d)^2 + 4bc}}{2}$$

These expressions for λ_1, λ_2 make no sense if $2 = 0$ (such as in the binary field). However, even though our field is assumed to satisfy $2 \neq 0$, a problem arises if $(a - d)^2 + 4bc$ is not the square of any scalar. We have three possible situations:

1. No roots: If there is no scalar $s \in \mathbb{F}$ for which $s^2 = (a - d)^2 + 4bc$, then there are no eigenvalues of A.
2. One (repeated) root: Suppose $(a - d)^2 + 4bc = 0$. Then there is one root, often called a "repeated root":
$$\lambda_1 = \lambda_2 = \frac{a + d}{2}$$

3. Two roots: Suppose there is a scalar $s \neq 0$ such that $s^2 = (a-d)^2 + 4bc$. Then we have two distinct roots:

$$\lambda_1 = \frac{(a+d)+s}{2}$$
$$\lambda_2 = \frac{(a+d)-s}{2}$$

Notice that if we have at least one root (so λ_1, λ_2 exist) then we have:

$$\lambda_1 + \lambda_2 = a + d = \text{trace}(A)$$
$$\lambda_1 \lambda_2 = ad - bc = \det(A)$$

So the sum of eigenvalues is the trace of the matrix, and the product of eigenvalues is the determinant. It can be shown this is true more generally for $n \times n$ matrices in the field $\mathbb{F} = \mathbb{C}$ and when eigenvalues are counted according to their algebraic multiplicity with respect to the polynomial equation $\det(A - \lambda I) = 0$ (see Section 12.8.3). The complex field has the nice property that all scalars are squares of some other scalar, which means every quadratic equation has at least one root. There are no finite fields for which all quadratic equations have at least one root (see Exercise 11.9).

For example, consider the field $\mathbb{F} = \mathbb{R}$ and the matrix A that corresponds to rotation by 90 degrees:

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

Intuitively, we know this has no eigenvectors because every input vector is rotated. The eigenvalue equation gives

$$\lambda = \frac{(a+d) \pm \sqrt{(a-d)^2 + 4bc}}{2} = \lambda = \frac{2 \pm \sqrt{-4}}{2}$$

which has no solution because, in the real field, -4 has no square root. However, if we consider this *same* matrix A in the complex field $\mathbb{F} = \mathbb{C}$ then we have two distinct eigenvalues

$$\lambda_1 = i \quad , \quad \lambda_2 = -i$$

which means that the corresponding eigenvectors v_1, v_2 are linearly independent. We can solve for the eigenvectors simultaneously: Let λ be an eigenvalue (either $\lambda = \lambda_1$ or $\lambda = \lambda_2$). We find a nontrivial vector in the null space of $A - \lambda I$. By Gaussian elimination we have:

$$\begin{bmatrix} -\lambda & -1 \\ 1 & -\lambda \end{bmatrix} \to \begin{bmatrix} 1 & -\lambda \\ -\lambda & -1 \end{bmatrix} \to \begin{bmatrix} 1 & -\lambda \\ 0 & 0 \end{bmatrix}$$

where the last elimination step holds because we already know the second row is a scalar multiple of the first row (since λ is an eigenvalue so the null space must be nontrivial). So we have a general eigenvector (a, b) with $a - \lambda b = 0$. So let $b = 1$ and $a = \lambda$. Then

$$\text{eigenvalue } \lambda \text{ has eigenvector} \quad v = (\lambda, 1)$$

and so $\lambda_1 = i$ has eigenvector $v_1 = (i, 1)$; $\lambda_2 = -i$ has eigenvector $v_2 = (-i, 1)$. We can collect this information into matrices Λ and S:

$$\Lambda = \begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix} \quad , \quad S = \begin{bmatrix} i & -i \\ 1 & 1 \end{bmatrix}$$

and one can verify that A has the diagonalizable form $A = S\Lambda S^{-1}$.

As another example, consider $\mathbb{F} = \mathbb{F}_7 = \{0, 1, 2, 3, 4, 5, 6\}$ with mod 7 arithmetic. This field has $1 \neq -1$ (since $-1 = 6$), *but not all scalars have square roots in this field.* As an exercise, one can find scalars $a, b, c, d \in \mathbb{F}_7$ such that the matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

has (i) No eigenvalues; (ii) Exactly one eigenvalue; (iii) Two distinct eigenvalues.

12.7. Diagonalizable matrices

Let A be a $n \times n$ matrix with entries in some field $\mathbb{F}$. We say that A is *diagonalizable* if there are square matrices S, Λ of the same size as A, where S is invertible and Λ is diagonal, such that $A = S\Lambda S^{-1}$. We can write S and Λ as

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \quad , \quad S = \begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & \cdots & | \end{bmatrix}$$

where $\lambda_1, \ldots, \lambda_n$ are the diagonal entries of Λ, and $v_1, \ldots, v_n$ are the columns of S.

If A is diagonalizable, so $A = S\Lambda S^{-1}$, we claim that: (i) $\{v_1, \ldots, v_n\}$ is a basis for $\mathbb{F}^n$; (ii) For each $i \in \{1, \ldots, n\}$, v_i is an eigenvector of A with corresponding eigenvalue λ_i. To see this, first note that S is invertible so its columns must be linearly independent. Thus, $\{v_1, \ldots, v_n\}$ are linearly independent vectors, there are n of them, so they are a basis for

$\mathbb{F}^n$. It follows that $v_i \neq 0$ for all i (a requirement of an eigenvector). Also

$$\underbrace{S^{-1}\begin{bmatrix} | & | & \cdots & | \\ v_1 & v_2 & \cdots & v_n \\ | & | & \cdots & | \end{bmatrix}}_{S} = \underbrace{\begin{bmatrix} | & | & \cdots & | \\ e_1 & e_2 & \cdots & e_n \\ | & | & \cdots & | \end{bmatrix}}_{I} \iff \left(S^{-1}v_i = e_i \quad \forall i \in \{1, \ldots, n\}\right)$$

Thus, for each $i \in \{1, \ldots, n\}$ we have

$$Av_i \overset{(a)}{=} S\Lambda S^{-1}v_i \overset{(b)}{=} S\Lambda e_i = S(\lambda_i e_i) = \lambda_i v_i$$

where (a) substitutes $A = S\Lambda S^{-1}$; (b) substitutes $S^{-1}v_i = e_i$; the final two equalities use the fact that Λe_i and Se_i is the ith column of Λ and S, respectively. So v_i is an eigenvector of A with eigenvalue λ_i.

Conversely, suppose we are given vectors $\{v_1, \ldots, v_n\}$ that are a basis for $\mathbb{F}^n$ and that are eigenvectors of A with eigenvalues λ_i. Then, we can define S as the matrix with columns $\{v_1, \ldots, v_n\}$ (so S is invertible), we can define Λ as the diagonal matrix of eigenvalues $\lambda_1, \ldots, \lambda_n$, and

$$(Av_i = \lambda_i v_i \quad \forall i) \iff (AS = S\Lambda)$$

Since S is invertible, the equation $AS = S\Lambda$ immediately implies $A = S\Lambda S^{-1}$, so A is diagonalizable. These findings are summarized in the following lemma.

LEMMA 12.5. *A $n \times n$ matrix A is diagonalizable if and only if it has eigenvectors $\{v_1, \ldots, v_n\}$ that are a basis for $\mathbb{F}^n$. (The corresponding eigenvalues $\lambda_1, \ldots, \lambda_n$ are possibly repeated.)*

To *diagonalize a matrix* means to find a basis of eigenvectors (if one exists), collect them into a matrix S (which is necessarily invertible), collect the corresponding (possibly repeated) eigenvalues into a matrix Λ, and then we immediately have $A = S\Lambda S^{-1}$.

12.7.1. Are there any more eigenvalues of A?

If a $n \times n$ matrix A is diagonalizable, so $A = S\Lambda S^{-1}$, we know $\{\lambda_1, \ldots, \lambda_n\}$ is a list of (possibly repeated) eigenvalues of A. Are there any other eigenvalues of A? The answer is "no," but this is not obvious. If the list $\{\lambda_1, \ldots, \lambda_n\}$ contains *n distinct* scalars then there can be no more eigenvalues, since we already know A has at most n eigenvalues. What if our list $\{\lambda_1, \ldots, \lambda_n\}$ has some repetitions? This leaves open the possibility that there is some other scalar $\lambda \notin \{\lambda_1, \ldots, \lambda_n\}$ that is an eigenvalue of A. Exercise 12.39 shows this is impossible.

12.7.2. Powers of a diagonalizable matrix

It is easy to compute powers of a diagonalizable matrix: If $A = S\Lambda S^{-1}$ then

$$\begin{aligned} A^2 &= (S\Lambda S^{-1})(S\Lambda S^{-1}) = S\Lambda(S^{-1}S)\Lambda S^{-1} = S\Lambda^2 S^{-1} \\ A^3 &= (A^2)A = S\Lambda S^{-1}S\Lambda S^{-1} = S\Lambda^3 S^{-1} \\ &\cdots \\ A^k &= S\Lambda^k S^{-1} \end{aligned}$$

If A is $n \times n$, then for any positive integer k, we have

$$\Lambda^k = \begin{bmatrix} \lambda_1^k & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^k & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3^k & \cdots & 0 \\ 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \cdots & \lambda_n^k \end{bmatrix}$$

12.7.3. A nondiagonalizable matrix

The simplest example of a matrix that is *not diagonalizable* is (using any field $\mathbb{F}$):

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

The eigenvalues are $\lambda_1 = \lambda_2 = 1$ (a repeated root). While $(1, 0)$ is clearly an eigenvector, there are no other eigenvectors that are not multiples of $(1, 0)$. We cannot find a basis of eigenvectors, so A is not diagonalizable. (See related Exercises 12.31 and 12.32.)

12.7.4. Examples: Diagonalizing a matrix

1. (Diagonalizing a matrix with real eigenvalues) Consider the real matrix

$$A = \begin{bmatrix} 3 & 3 \\ 2 & 1 \end{bmatrix} \tag{12.8}$$

The eigenvalue equation is

$$\det(A - \lambda I) = (3 - \lambda)(1 - \lambda) - 6 = 0$$

Solving gives

$$\lambda_1 = 2 + \sqrt{7} \quad , \quad \lambda_2 = 2 - \sqrt{7} \tag{12.9}$$

To find the eigenvector for λ (where $\lambda = \lambda_1$ or $\lambda = \lambda_2$) we look for vectors in $Null(A - \lambda I)$. We already know this null space is nontrivial. By Gaussian elimination

$$\begin{bmatrix} 3-\lambda & 3 \\ 2 & 1-\lambda \end{bmatrix} \to \begin{bmatrix} 3-\lambda & 3 \\ 0 & 0 \end{bmatrix}$$

where the Gaussian elimination step multiplies row 1 by $-2(3-\lambda)^{-1}$ and adds it into row 2. (We already know $3 - \lambda \neq 0$, and also we already know row 2 must be a multiple of row 1, else it would introduce another pivot and the null space would be trivial.) So for eigenvector (x, y) we get

$$(3-\lambda)x + 3y = 0$$

Consider the form $(x, y) = (1, y)$. Then $y = (1/3)\lambda - 1$. So we obtain eigenvectors

$$v_1 = \begin{bmatrix} 1 \\ (1/3)\lambda_1 - 1 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 1 \\ (1/3)\lambda_2 - 1 \end{bmatrix}$$

Collecting this info gives

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} 2+\sqrt{7} & 0 \\ 0 & 2-\sqrt{7} \end{bmatrix}$$

$$S = \begin{bmatrix} 1 & 1 \\ (1/3)\lambda_1 - 1 & (1/3)\lambda_2 - 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{-1+\sqrt{7}}{3} & \frac{-1-\sqrt{7}}{3} \end{bmatrix}$$

We can verify the diagonalization: $A = S\Lambda S^{-1}$ is correct.

2. (Diagonalizing a matrix with complex eigenvalues) Consider the real matrix

$$A = \begin{bmatrix} -1 & -5 \\ 3 & 2 \end{bmatrix} \tag{12.10}$$

The eigenvalue equation is

$$\det(A - \lambda I) = (-1-\lambda)(2-\lambda) + 15 = 0$$

There are no roots over the field $\mathbb{R}$. *For these problems with real matrices, unless stated*

otherwise, we will always allow ourselves to expand into the complex field. We treat A as a complex-valued matrix using field $\mathbb{F} = \mathbb{C}$ and we obtain two distinct eigenvalues:

$$\lambda_1 = \frac{1 + i\sqrt{51}}{2} \quad , \quad \lambda_2 = \frac{1 - i\sqrt{51}}{2} \tag{12.11}$$

To find the eigenvector for λ (where $\lambda = \lambda_1$ or $\lambda = \lambda_2$) we look for vectors in $Null(A - \lambda I)$. We already know this null space is nontrivial. By Gaussian elimination

$$\begin{bmatrix} -1-\lambda & -5 \\ 3 & 2-\lambda \end{bmatrix} \to \begin{bmatrix} -1-\lambda & -5 \\ 0 & 0 \end{bmatrix}$$

where the Gaussian elimination step multiplies row 1 by $-3(-1-\lambda)^{-1}$ and adds it into row 2. (We already know $-1-\lambda \neq 0$, and also we already know row 2 must be a multiple of row 1, else it would introduce another pivot and the null space would be trivial.) So for eigenvector (x, y) we get

$$(-1-\lambda)x - 5y = 0$$

Consider the form $(x, y) = (1, y)$. Then $y = (-1/5)(\lambda + 1)$. So we obtain eigenvectors

$$v_1 = \begin{bmatrix} 1 \\ (-1/5)(\lambda_1 + 1) \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 1 \\ (-1/5)(\lambda_2 + 1) \end{bmatrix}$$

Collecting this info gives

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix} = \begin{bmatrix} \frac{1+i\sqrt{51}}{2} & 0 \\ 0 & \frac{1-i\sqrt{51}}{2} \end{bmatrix}$$

$$S = \begin{bmatrix} 1 & 1 \\ (-1/5)(\lambda_1 + 1) & (-1/5)(\lambda_2 + 1) \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ \frac{-3-i\sqrt{51}}{10} & \frac{-3+i\sqrt{51}}{10} \end{bmatrix}$$

We verify the diagonalization (say, by computer): $A = S\Lambda S^{-1}$ is correct.

3. (Diagonalizing a matrix with field $\mathbb{F}_5$) Consider $\mathbb{F}_5 = \{0, 1, 2, 3, 4\}$ with mod 5 arithmetic. Define the following matrix with entries in this field

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

The matrix is upper triangular so its eigenvalues are 1 and 2. Solving for vectors (x, y)

in $Null(A - \lambda I)$, for $\lambda = 1$ or $\lambda = 2$, gives

$$(1 - \lambda)x + 4y = 0$$
$$(2 - \lambda)y = 0$$

For $\lambda = 1$ we must have $y = 0$ (so choose $x = 1$); for $\lambda = 2$ we can choose $y = 1$ and $x = 4$. Then

$$\lambda_1 = 1, v_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix} \quad ; \quad \lambda_2 = 2, v_2 = \begin{bmatrix} 4 \\ 1 \end{bmatrix}$$

Collecting this info

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix} \quad , \quad S = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$$

We can compute S^{-1} using the formula for a 2×2:

$$S^{-1} = 1^{-1} \begin{bmatrix} 1 & -4 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

We can verify the diagonalization: $A = S\Lambda S^{-1}$ is correct.

12.8. Similar matrices

Fix $\mathbb{F}$ as a field. Let n be a positive integer. Two $n \times n$ matrices A, B are said to be *similar* if

$$A = PBP^{-1}$$

for some invertible $n \times n$ matrix P. It is easy to show that if A, B are similar, then B, A are also similar. We can use notation $A \sim B$ to mean that matrices A and B are similar.

12.8.1. Consequences of similarity

1. A square matrix A is diagonalizable if and only if it is similar to a diagonal matrix.
2. Similar matrices have the same determinant.
3. Similar matrices have the same trace.
4. Similar matrices have the same eigenvalues.

A quick test to see if two matrices are similar is to compute their traces: If they have different traces, they cannot be similar. If they have the same trace, the test is inconclusive.

For example, in the real field:

$$A = \begin{bmatrix} 1 & 7 \\ 2 & 2 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 4 & 1 \\ -1 & 4 \end{bmatrix}$$

Since $\text{trace}(A) = 3$ and $\text{trace}(B) = 8$, matrices A and B are not similar.

- Proof of Claim 1: Suppose A is diagonalizable. Then $A = S\Lambda S^{-1}$ where Λ is a diagonal matrix of eigenvalues. So A is similar to the diagonal matrix Λ. Conversely, suppose A is similar to a diagonal matrix D. Then $A = PDP^{-1}$ for some invertible matrix P, which means A is diagonalizable (the diagonal entries of D are the eigenvalues of A, the columns of P are the corresponding eigenvectors).
- Proof of Claim 2: Suppose $A = PBP^{-1}$. By properties of determinants:

$$\det(A) = \det(P)\det(B)\det(P^{-1}) = \det(B)\det(P)\det(P^{-1}) = \det(B)\underbrace{\det(PP^{-1})}_{\det(I)} = \det(B)$$

- Proof of Claim 3: Suppose $A = PBP^{-1}$. Then

$$\text{trace}(A) = \text{trace}(PBP^{-1}) = \text{trace}(P(BP^{-1})) \overset{(a)}{=} \text{trace}((BP^{-1})P) = \text{trace}(BI) = \text{trace}(B)$$

  where equality (a) uses the property of trace given in (2.16).
- Proof of Claim 4: Suppose $A = PBP^{-1}$. Let λ be an eigenvalue of A.We now show that λ is also an eigenvalue of B. Since λ is an eigenvalue of A, there is a vector $v \neq 0$ such that $Av = \lambda v$. Thus

$$\left(\underbrace{(PBP^{-1})}_{A} v = \lambda v \right) \implies B(P^{-1}v) = \lambda(P^{-1}v)$$

  where the implication follows by multiplying both sides by P^{-1}. Since the vector $P^{-1}v$ is nonzero (why?), it follows that λ is an eigenvalue of B with eigenvector $P^{-1}v$. This argument shows that any eigenvalue of A is also an eigenvalue of B. Swapping the roles of A and B implies that any eigenvalue of B is also an eigenvalue of A.

12.8.2. Similarity and change of basis

If A and B are similar, they represent matrices for the *same* linear transformation under different bases. Specifically, let A be a square matrix with entries in $\mathbb{F}$. Define the linear

transformation $T : \mathbb{F}^n \to \mathbb{F}^n$ by

$$T(x) = Ax \quad \forall x \in \mathbb{F}^n$$

Let $\{e_1, ..., e_n\}$ be the standard basis for $\mathbb{F}^n$. Then every input $x = (x_1, ..., x_n) \in \mathbb{F}^n$ and every output $y = (y_1, ..., y_n) = Ax$ can be uniquely expressed in terms of this basis:

$$x = \sum_{i=1}^{n} x_i e_i$$
$$y = \sum_{i=1}^{n} y_i e_i$$

Every input coefficient vector $x \in \mathbb{F}^n$ has output coefficient vector $y = Ax$. Therefore, the matrix $Mat(T)$ associated with the linear transformation T under the basis $\{e_1, ..., e_n\}$ (used for both input and output space) is just A itself:

$$Mat(T) = A$$

Now let P be an invertible matrix. Consider the following alternative basis for $\mathbb{F}^n$:

$$\{Pe_1, Pe_2, ..., Pe_n\}$$

That is, the columns of P form a basis for $\mathbb{F}^n$. Every input $x = (x_1, ..., x_n) \in \mathbb{F}^n$ and every output $y = (y_1, ..., y_n) \in \mathbb{F}^n$ can be uniquely expressed as linear combinations of these new basis vectors using some scalar coefficients u_i and v_i, respectively:

$$x = \sum_{i=1}^{n} u_i Pe_i$$
$$y = \sum_{i=1}^{n} v_i Pe_i$$

The above equations are the same as matrix multiplication:

$$x = Pu$$
$$y = Pv$$

where $u = (u_1, ..., u_n)$ and $v = (v_1, ..., v_n)$. In particular, given any $u \in \mathbb{F}^n$ we can obtain x via $x = Pu$; given any $x \in \mathbb{F}^n$ we can obtain u via $u = P^{-1}x$. Similarly, we can switch between y and v by multiplying by either P or P^{-1}.

Let B denote $Mat(T)$ under this new basis $\{Pe_1, ..., Pe_n\}$. Then B represents the linear

relationship between the input coefficients u and the output coefficients v:

$$v = Bu$$

We can explicitly solve for B as follows: Using $v = P^{-1}y$ and $u = P^{-1}x$ in the above equality gives

$$P^{-1}y = BP^{-1}x$$

Multiplying both sides by P gives

$$y = PBP^{-1}x$$

This holds for all $x \in \mathbb{F}^n$. In particular, we have

$$y = PBP^{-1}x \quad \forall x \in \mathbb{F}^n$$
$$y = Ax \quad \forall x \in \mathbb{F}^n$$

Therefore, A and PBP^{-1} must be the same matrices:

$$A = PBP^{-1}$$

which means $B = P^{-1}AP$. The matrices A and B are similar. The matrix A expresses the linear transformation $T(x) = Ax$ under the standard basis, and the matrix B expresses this same linear transformation under a new basis.

12.8.3. Similarity to an upper triangular matrix

Consider the complex field $\mathbb{C}$. Since every real number can be viewed as a special case of a complex number, every real matrix can be viewed as a special case of a matrix with complex-valued entries. The complex field admits the following result.

THEOREM 12.2. *Let A be a square matrix with entries in $\mathbb{C}$. Then*

a) Matrix A has an eigenvalue $\lambda \in \mathbb{C}$.

b) Matrix A is similar to an upper triangular matrix U. That is, there are matrices U and P of the same size as A (and with entries in $\mathbb{C}$), where U is upper triangular and P is invertible, such that $A = PUP^{-1}$.

c) For every decomposition $A = PUP^{-1}$ where U is upper triangular and P is invertible, the diagonal entries of U are exactly the eigenvalues of A. Also, regardless of which P and U matrices are used in the decomposition $A = PUP^{-1}$, each distinct eigenvalue of A is always repeated the same number of times on the diagonal of U. The number of times an eigenvalue of A is repeated on the diagonal of U is called the algebraic multiplicity *of that eigenvalue.*

The proof of (a) uses the *fundamental theorem of algebra* to ensure there is a $\lambda \in \mathbb{C}$ such

that $\det(A - \lambda I) = 0$ (see Chapter 13). Chapter 13 also relates this definition of *algebraic multiplicity* to the multiplicity of roots for the equation $\det(A - \lambda I) = 0$. The proof of (b) and (c) uses a deep theory of *invariant subspaces* that is discussed in Chapter 13.

Theorem 12.2 guarantees that, using the complex field $\mathbb{C}$, every $n \times n$ matrix A has k distinct eigenvalues $\lambda_1, \ldots, \lambda_k \in \mathbb{C}$, where $1 \leq k \leq n$. Let m_i be the algebraic multiplicity of eigenvalue λ_i. Since U has n diagonal entries, and each eigenvalue of A appears on the diagonal of U a number of times equal to its algebraic multiplicity, we have

$$m_1 + m_2 + \ldots + m_k = n$$

It is common to relabel the list of k eigenvalues to a list $\{\lambda_1, \ldots, \lambda_n\}$ *that always has exactly n eigenvalues,* each repeated according to algebraic multiplicity.

COROLLARY 12.1. *Let A be a $n \times n$ matrix with real or complex entries. Using field $\mathbb{C}$, let $\{\lambda_1, \ldots, \lambda_n\}$ be the list of n (possibly complex) eigenvalues of A (each distinct eigenvalue being redundantly listed a number of times equal to its algebraic multiplicity). Then*

a) $\det(A) = \prod_{i=1}^{n} \lambda_i$.

b) $\mathit{trace}(A) = \sum_{i=1}^{n} \lambda_i$.

PROOF. By Theorem 12.2, we know $A = PUP^{-1}$ for some invertible P and some upper triangular U. Since A and U are similar, we know $\det(A) = \det(U)$ and $\mathrm{trace}(A) = \mathrm{trace}(U)$. Also, the diagonal entries of the upper triangular matrix U are $\{\lambda_1, \ldots, \lambda_n\}$ and so

$$\det(U) = \prod_{i=1}^{n} \lambda_i$$

$$\mathrm{trace}(U) = \sum_{i=1}^{n} \lambda_i$$

□

Notes:

- If a $n \times n$ matrix A has purely real entries, its trace and determinant must both be real. However, one or more of its eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ might be complex. Therefore, the eigenvalues (repeated according to multiplicity) must have a structure that ensures their sum and product are both purely real. For example, the integer matrix A in (12.10) has eigenvalues $\lambda_1 = (1 + i\sqrt{51})/2$ and $\lambda_2 = (1 - i\sqrt{51})/2$ (as computed in (12.11)), and we have $\lambda_1 + \lambda_2 = 1$ and $\lambda_1 \lambda_2 = 13$.
- If a $n \times n$ matrix A has all integer entries, its trace and determinant must be integers, even though its eigenvalues $\{\lambda_1, \ldots, \lambda_n\}$ are not necessarily integer. Thus, the (possibly noninteger) eigenvalues must have an integer sum and an integer product. For example, the integer matrix A in (12.8) has eigenvalues $\lambda_1 = 2 + \sqrt{7}$ and $\lambda_2 = 2 - \sqrt{7}$ (as computed

in (12.9)), and we have $\lambda_1 + \lambda_2 = 4$ and $\lambda_1\lambda_2 = -3$.

- If we compute $n - 1$ eigenvalues of a $n \times n$ matrix (with real or complex entries), we can use the trace formula as an easy way to get the final eigenvalue.
- Corollary 12.1 holds more generally for square matrices A with entries in a general field $\mathbb{F}$, provided that A is similar to an upper triangular matrix U. Then, the determinant is the product of eigenvalues of A (being the diagonal entries of U, with multiplicity defined by the number of repetitions on this diagonal); the trace is the sum of eigenvalues. Of course, with a general field $\mathbb{F}$, we cannot guarantee that A is similar to an upper triangular matrix. On the other hand, any matrix A that is diagonalizable is similar to a diagonal matrix, and diagonal matrices are upper triangular. Also, for any field $\mathbb{F}$, a $n \times n$ matrix with n *distinct* eigenvalues must be diagonalizable (since the corresponding n eigenvectors are linearly independent and hence are a basis for $\mathbb{F}^n$), so the sum and product of these n distinct eigenvalues is the trace and determinant.

12.9. Eigenvalues for linear operators

Let V be a general vector space over a field $\mathbb{F}$. The concept of eigenvalue and eigenvector can be extended to linear transformations $T : V \to V$. Recall that a linear transformation from a vector space to that same vector space is also called a *linear operator*. A scalar $\lambda \in \mathbb{F}$ is defined to be an *eigenvalue of* T if

$$T(v) = \lambda v$$

for some nonzero vector $v \in V$. The vector v is called an *eigenvector of* T. The pair (λ, v) is called an *eigenpair of* T.

The result of Theorem 12.1 can be extended to this context: If $\lambda_1, \ldots, \lambda_k$ are distinct eigenvalues of a linear operator $T : V \to V$, then the corresponding eigenvectors $\{v_1, \ldots, v_k\}$ are linearly independent. It follows that the number of eigenvalues is less than or equal to $\dim(V)$. In the trivial case $\dim(V) = 0$, then $V = \{0\}$ and there are no eigenvectors because V has no nonzero vectors. If $\dim(V) = n$ for some positive integer n, then T has at most n eigenvalues. If $\dim(V) = \infty$, then T can have an infinite number of eigenvalues.

For example, let V be the real vector space of infinitely differentiable functions $f : \mathbb{R} \to \mathbb{R}$. This vector space has infinite dimension because it contains all polynomial functions. Define the linear operator $T : V \to V$ for all $f \in V$ by the derivative map: $T(f) = f'(t)$ for all $t \in \mathbb{R}$. Remarkably, each $\lambda \in \mathbb{R}$ is an eigenvalue of T. This is because, given $\lambda \in \mathbb{R}$, we can define a function $f_\lambda \in V$ by

$$f_\lambda(t) = e^{\lambda t} \quad \forall t \in \mathbb{R}$$

and we observe that f_λ is an eigenvector with eigenvalue λ because

$$\begin{aligned} T(f_\lambda) &= f'_\lambda(t) \quad \forall t \in \mathbb{R} \\ &= \lambda e^{\lambda t} \quad \forall t \in \mathbb{R} \\ &= \lambda f_\lambda \end{aligned}$$

In this vector space of functions, the eigenvector f_λ is often called an *eigenfunction* because it is a function of t.

As an exercise, the reader is invited to consider the real vector space V of all functions $f : \mathbb{R} \to \mathbb{R}$. Define the linear operator $T : V \to V$ for all $f \in V$ by

$$T(f) = \cos(t) f(t) \quad \forall t \in \mathbb{R}$$

(a) Show that if $\lambda \notin [-1, 1]$ then λ is not an eigenvalue of T. Here, $[-1, 1]$ denotes the interval $\{x \in \mathbb{R} : -1 \leq x \leq 1\}$.

(b) Show that every $\lambda \in [-1, 1]$ is an eigenvalue of T. What are the corresponding eigenfunctions f_λ?

(c) Define the subspace $S \subseteq V$ as the set of all *continuous functions* in V. Let $G : S \to S$ be the linear operator that restricts T to S. That is, for each $f \in S$ define $G(f) = \cos(t) f(t)$ for all $t \in \mathbb{R}$. Show that G has no eigenvalues.

The *spectrum of a linear operator* $T : V \to V$, in the special case $\dim(V) < \infty$, is defined as the set of eigenvalues of T. This is denoted $\Lambda(T)$ and is a finite and possibly empty set. When $\dim(V) = \infty$, the definition of *spectrum* is often loaded with additional requirements. In this text, we do not define the term *spectrum* in the infinite-dimensional context.

12.10. Exercises

EXERCISE 12.1. *Suppose A is a square matrix with eigenpair (λ, v). Find an eigenpair of A^2.*

EXERCISE 12.2. *Prove a square matrix A is invertible if and only if* 0 *is* not *an eigenvalue of A.*

EXERCISE 12.3. *Find all eigenvalues of the following real matrix:*

$$A = \begin{bmatrix} 5 & 5 & 5 & 5 \\ 2 & 2 & 2 & 2 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

EXERCISE 12.4. *Find all eigenvalues of the following real matrix:*

$$A = \begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix}$$

EXERCISE 12.5. *Define the following real matrix:*

$$A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix}$$

a) Find two eigenvalues λ_1, λ_2 and the corresponding eigenvectors v_1, v_2.

b) Use a computer to verify $A = S\Lambda S^{-1}$ for $S = \begin{bmatrix} v_1 & v_2 \end{bmatrix}$ and

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$$

EXERCISE 12.6. *Define the following real matrix*

$$A = \begin{bmatrix} 2 & 2 \\ 0 & 1 \end{bmatrix}$$

a) Find all eigenvalues of A.

b) Find all eigenvalues of $A^\top$.

EXERCISE 12.7. *Define the following real matrix*

$$A = \begin{bmatrix} 2 & 2 \\ 3 & 1 \end{bmatrix}$$

a) Find all eigenvalues of A.

b) Find all eigenvalues of $A^\top$.

EXERCISE 12.8. *Define the following real matrix*

$$A = \begin{bmatrix} 5 & -2 \\ 0 & 1 \end{bmatrix}$$

a) Find all eigenvalues of A.

b) Find all eigenvalues of $A^\top$.

EXERCISE 12.9. *Define the following real matrix*

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix}$$

a) Find all eigenvalues of A.

b) Find all eigenvalues of $A^\top$.

EXERCISE 12.10. *Define the following real matrix*

$$A = \begin{bmatrix} 2 & 4 \\ 1 & 2 \end{bmatrix}$$

a) Find all eigenvalues of A.

b) Find all eigenvalues of $A^\top$.

EXERCISE 12.11. *Define the following real matrix*

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}$$

a) Find all eigenvalues of A.

b) Find all eigenvalues of $A^\top$.

EXERCISE 12.12. *(Eigenvalue of $A^\top$) Let A be a square matrix with entries in some field $\mathbb{F}$. Suppose A has an eigenvalue $\lambda \in \mathbb{F}$. Use the fact $\det(M) = \det(M^\top)$ to prove that λ is also an eigenvalue of $A^\top$.*

EXERCISE 12.13. *Diagonalize the following real matrix A:*

$$A = \begin{bmatrix} 4 & 4 \\ 1 & 2 \end{bmatrix}$$

EXERCISE 12.14. *(Matrix with complex eigenvalues) Diagonalize the following real matrix A:*

$$A = \begin{bmatrix} 2 & 4 \\ -1 & 3 \end{bmatrix}$$

EXERCISE 12.15. *(Matrix with complex eigenvalues) Diagonalize the following real matrix A:*

$$A = \begin{bmatrix} 10 & -5 \\ 30 & 10 \end{bmatrix}$$

EXERCISE 12.16. *Diagonalize this real matrix A*

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

EXERCISE 12.17. *Diagonalize this real matrix A*

$$A = \begin{bmatrix} 7 & -1 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}$$

EXERCISE 12.18. *Diagonalize this real matrix A*

$$A = \begin{bmatrix} 10 & 0 & 0 \\ 0 & 5 & 1 \\ 0 & 2 & 1 \end{bmatrix}$$

EXERCISE 12.19. *(Field $\mathbb{F}_3$) Diagonalize A using $\mathbb{F}_3 = \{0, 1, 2\}$ and mod 3 arithmetic:*

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 2 \end{bmatrix}$$

EXERCISE 12.20. *(Field $\mathbb{F}_3$) Diagonalize A using $\mathbb{F}_3 = \{0, 1, 2\}$ and mod 3 arithmetic:*

$$A = \begin{bmatrix} 2 & 1 & 1 \\ 0 & 0 & 2 \\ 0 & 0 & 1 \end{bmatrix}$$

EXERCISE 12.21. *(Field $\mathbb{F}_2$) Does the following matrix have any eigenvalues in $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic?*

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

EXERCISE 12.22. *(Field $\mathbb{F}_3$) Consider matrix A with $\mathbb{F}_3 = \{0, 1, 2\}$ with mod 3 arithmetic:*

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

a) Write the polynomial $f(\lambda) = \det(A - \lambda I) = 0$.

b) Does $f(\lambda)$ have any solutions $\lambda \in \{0, 1, 2\}$?

EXERCISE 12.23. *Consider the real matrix A:*

$$A = \begin{bmatrix} 9 & 4 \\ 3 & 2 \end{bmatrix}$$

a) Find the degree 2 polynomial $f(\lambda) = \det(A - \lambda I)$. Solve to find λ_1, λ_2.

b) Find $\sum_{i=1}^{2} \lambda_i$ and $\prod_{i=1}^{2} \lambda_i$

c) Find all eigenvalues of $100A$.

EXERCISE 12.24. *Consider the real matrix A:*

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

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ and $\prod_{i=1}^{3} \lambda_i$

EXERCISE 12.25. *Consider the real matrix A:*

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

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ and $\prod_{i=1}^{3} \lambda_i$

EXERCISE 12.26. *Consider the real matrix A:*

$$A = \begin{bmatrix} 3 & 1 & 9 \\ 2 & 1 & 2 \\ 0 & 4 & 4 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

EXERCISE 12.27. *Consider the real matrix A:*

$$A = \begin{bmatrix} 2 & 2 & 2 \\ 4 & 0 & 2 \\ -1 & 0 & 0 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

EXERCISE 12.28. *(Complex eigenvalues) Consider the real matrix A:*

$$A = \begin{bmatrix} 2 & -2 & -1 \\ 4 & 0 & 2 \\ -1 & 0 & 0 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

EXERCISE 12.29. *(Repeating a column) Consider the real matrix A:*

$$A = \begin{bmatrix} 1 & 1 & 2 \\ 4 & 4 & 2 \\ -1 & -1 & 0 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$. *Find* $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

c) Find all eigenvalues of $-7A$.

EXERCISE 12.30. *(Repeating a row) Consider the real matrix A:*

$$A = \begin{bmatrix} 9 & 9 & 4 \\ 9 & 9 & 4 \\ 3 & -1 & 0 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$. *Find* $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

c) Find all eigenvalues of $4A$.

EXERCISE 12.31. *(Cannot be diagonalized) Show this real matrix cannot be diagonalized:*

$$A = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}$$

EXERCISE 12.32. *(Diagonalizable depending on c) Show that the following real matrix can be diagonalized if and only if* $c = 0$*:*

$$A = \begin{bmatrix} 3 & c \\ 0 & 3 \end{bmatrix}$$

EXERCISE 12.33. *Consider* $\mathbb{F} = \mathbb{R}$. *Let* P *be a* 2×2 *invertible matrix. Define the real matrix*

$$A = \begin{bmatrix} 10 & -1 \\ 2 & 2 \end{bmatrix}$$

a) Find all eigenvalues of PAP^{-1} *in terms of* P.

b) Diagonalize PAP^{-1} *in terms of* P.

EXERCISE 12.34. *Consider* $\mathbb{F} = \mathbb{R}$. *Let* P *be a* 2×2 *invertible matrix. Define the real matrix*

$$A = \begin{bmatrix} 1 & 4 \\ 8 & 2 \end{bmatrix}$$

a) Find all eigenvalues of PAP^{-1} *in terms of* P.

b) Diagonalize PAP^{-1} *in terms of* P.

EXERCISE 12.35. *Consider* $\mathbb{F} = \mathbb{R}$. *Let* P *be a* 3×3 *invertible matrix. Define the real matrix*

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 3 & 3 & 0 \\ 0 & 0 & 9 \end{bmatrix}$$

a) Find all eigenvalues of PAP^{-1} *in terms of* P.

b) Diagonalize PAP^{-1} *in terms of* P.

EXERCISE 12.36. *(Similar or not?) Are these two real matrices* A *and* B *similar?*

a)

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

b)

$$A = \begin{bmatrix} 1 & 9 & 2 \\ 0 & 4 & 3 \\ 0 & 0 & 5 \end{bmatrix}, B = \begin{bmatrix} 5 & 8 & 4 \\ 0 & 1 & 9 \\ 0 & 0 & 3 \end{bmatrix}$$

c)

$$A = \begin{bmatrix} 4 & 0 \\ 0 & 3 \end{bmatrix}, B = \begin{bmatrix} 3 & 1 \\ 0 & 4 \end{bmatrix}$$

EXERCISE 12.37. *(Similar or not?) Are these two real matrices A and B similar?*

a)

$$A = \begin{bmatrix} 1 & 4 \\ 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 3 & 9 \\ 1 & 2 \end{bmatrix}$$

b)

$$A = \begin{bmatrix} 1 & 3 & 1 \\ 5 & -4 & 5 \\ 7 & 8 & 7 \end{bmatrix}, B = \begin{bmatrix} 2 & 1 & 0 \\ 0 & 1 & -4 \\ 0 & 0 & 3 \end{bmatrix}$$

c)

$$A = \begin{bmatrix} 8 & 0 \\ 0 & 2 \end{bmatrix}, B = \begin{bmatrix} 3 & 5 \\ 1 & 7 \end{bmatrix}$$

EXERCISE 12.38. *Give your own example of a 2×2 matrix A in the real field that is not symmetric and that has two distinct eigenvalues. Compute the eigenvalues and corresponding eigenvectors for both A and $A^\top$. You should find the eigenvalues are the same, but the their corresponding eigenvectors cannot be the same.*

EXERCISE 12.39. *(Are there any other eigenvalues?) Let A be a diagonalizable $n \times n$ matrix, so $A = S\Lambda S^{-1}$ for some invertible S and some diagonal Λ. Let $\{\lambda_1, \ldots, \lambda_n\}$ be the diagonal elements of Λ (with possible repetitions) and let $\{v_1, \ldots, v_n\}$ be the columns of S. Suppose there is an eigenvalue λ of A such that $\lambda \notin \{\lambda_1, \ldots, \lambda_n\}$. We reach a contradiction.*

a) Let v be the corresponding eigenvector for λ. Explain why $v = \sum_{i=1}^n a_i v_i$ for some scalars a_i.

b) Multiply both sides of the equation $v = \sum_{i=1}^n a_i v_i$ by A.

c) Multiply both sides of the equation $v = \sum_{i=1}^n a_i v_i$ by λ. Reach a contradiction.

CHAPTER 13.

Polynomials, roots, and eigenvalues over $\mathbb{C}$

This chapter presents a factoring result that holds for polynomials over the field of complex numbers. The result implies that every square matrix with real or complex coefficients has a (possibly complex-valued) eigenvalue. It also relates the algebraic multiplicity of roots of a polynomial equation to the algebraic multiplicity of eigenvalues of a square matrix. The last part of this chapter defines an *invariant subspace* and shows that if A is a complex-valued square matrix with a nontrivial invariant subspace S, then there is an eigenvector of A in S. This fact is crucial for proving the *spectral theorem* for real and symmetric matrices (developed in Chapter 15).

13.1. Polynomials over the real field

A polynomial $p : \mathbb{R} \to \mathbb{R}$ is a function of the type

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots a_k x^k \quad \forall x \in \mathbb{R}$$

where k is a given nonnegative integer and $a_0, a_1, \ldots, a_k$ are given real numbers, called the *coefficients* of the polynomial. If k is a positive integer and $a_k \neq 0$, the polynomial is said to have *degree* k. A *constant polynomial* is a polynomial of the type

$$p(x) = a_0 \quad \forall x \in \mathbb{R}$$

for some coefficient $a_0 \in \mathbb{R}$. Constant polynomials are said to have *degree zero*. A polynomial is *nonconstant* if it is not a constant polynomial. It can be shown that nonconstant polynomials take an infinite number of different $p(x)$ values as x sweeps over the real number line. A polynomial is uniquely determined by its degree k and its coefficients $a_0, \ldots, a_k$. Two polynomial functions are the same function if and only if they have the same degree and the same coefficients. A value $r \in \mathbb{R}$ is called a *root* of a polynomial p if $p(r) = 0$.

The quadratic polynomial $p : \mathbb{R} \to \mathbb{R}$ defined by $p(x) = x^2 + 1$ for all $x \in \mathbb{R}$ clearly has no roots $r \in \mathbb{R}$. Now consider the polynomial $q : \mathbb{R} \to \mathbb{R}$ defined by

$$q(x) = x^3 + 2x^2 - 88x + 4 \quad \forall x \in \mathbb{R}$$

Limiting behavior can be understood by looking at the dominant term x^3:

$$\lim_{x \to \infty} q(x) = \infty$$
$$\lim_{x \to -\infty} q(x) = -\infty$$

Since q is a continuous function, the intermediate value theorem implies there must be a root $r \in \mathbb{R}$ such that $q(r) = 0$. By the same reasoning, every polynomial $p : \mathbb{R} \to \mathbb{R}$ with *odd degree* must have at least one root in the real number line. On the other hand, a polynomial $p : \mathbb{R} \to \mathbb{R}$ with *even degree* may or may not have a root in the real number line (consider, for example, the polynomial $p(x) = x^{2k} + 1$ for $k \in \{1, 2, 3, \ldots\}$).

13.2. Polynomials over the complex field

A polynomial $p : \mathbb{C} \to \mathbb{C}$ is a function of the type

$$p(x) = a_0 + a_1 x + a_2 x^2 + \cdots a_k x^k \quad \forall x \in \mathbb{C}$$

where k is a given nonnegative integer and $a_0, a_1, \ldots, a_k$ are given complex numbers called *coefficients*. The terminology of *degree, constant polynomial, nonconstant polynomial* is the same as for the case $\mathbb{R}$. A value $r \in \mathbb{C}$ is called a *root* if $p(r) = 0$.

Suppose p is a nonconstant polynomial. If k is a positive integer and $p : \mathbb{C} \to \mathbb{C}$ is a degree-k polynomial with root r, it can be shown that

$$p(x) = (x - r)q(x) \quad \forall x \in \mathbb{C} \tag{13.1}$$

for some degree-$(k-1)$ polynomial $q : \mathbb{C} \to \mathbb{C}$. Solving for the decomposition (13.1) when a root r is known is called *factoring out the term* $(x - r)$.

THEOREM 13.1. *(Polynomials with field $\mathbb{C}$) Let $p : \mathbb{C} \to \mathbb{C}$ be a polynomial.*

a) If p is a nonconstant polynomial, then there is at least one $r \in \mathbb{C}$ that satisfies $p(r) = 0$.

b) If p is a nonconstant polynomial with degree k, so that

$$p(x) = a_k x^k + a_{k-1} x^{k-1} + \ldots + a_1 x + a_0 \quad \forall x \in \mathbb{C}$$

for some coefficients $a_0, \ldots, a_k \in \mathbb{C}$ (with $a_k \neq 0$), then it can be factored as

$$p(x) = a_k (x - r_1)(x - r_2) \cdots (x - r_k) \quad \forall x \in \mathbb{C}$$

for some (possibly repeated) roots $r_1, \ldots, r_k \in \mathbb{C}$. In particular, $p(z) = 0$ if and only if $z \in \{r_1, r_2, \ldots, r_k\}$.

c) If all coefficients of p are real (so $a_i \in \mathbb{R}$ for all $i \in \{0, \ldots, k\}$) then all roots come in complex conjugate pairs [complex conjugates are discussed in Chapter 10.1]:

$$p(z) = 0 \iff p(\overline{z}) = 0$$

PROOF. Part (a) is the *fundamental theorem of algebra* (we omit a proof). Part (b) follows from part (a) together with the factoring result stated in (13.1). To prove part (c), suppose k is a positive integer. Let $p : \mathbb{C} \to \mathbb{C}$ be a degree-k polynomial of the form

$$p(x) = a_k x^k + a_{k-1} x^{k-1} + \ldots + a_1 x + a_0$$

for some real coefficients $a_k, a_{k-1}, \ldots, a_0$. Fix $z \in \mathbb{C}$ and suppose z is a root, so

$$0 = p(z)$$

Taking complex conjugates of both sides and using the fact $\overline{0} = 0$ gives

$$\begin{aligned}
0 &= \overline{p(z)} \\
&= \overline{a_k z^k + a_{k-1} z^{k-1} + \ldots + a_1 z + a_0} \\
&\overset{(a)}{=} \overline{a_k z^k} + \overline{a_{k-1} z^{k-1}} + \ldots + \overline{a_1 z} + \overline{a_0} \\
&\overset{(b)}{=} \overline{a_k} \cdot \overline{z}^k + \overline{a_{k-1}} \cdot \overline{z}^{k-1} + \ldots + \overline{a_1} \cdot \overline{z} + \overline{a_0} \\
&\overset{(c)}{=} a_k \cdot \overline{z}^k + a_{k-1} \cdot \overline{z}^{k-1} + \ldots + a_1 \cdot \overline{z} + a_0 \\
&= p(\overline{z})
\end{aligned}$$

where (a) holds because the conjugate of a sum is the sum of conjugates; (b) holds because the conjugate of a product is the product of conjugates; (c) holds because all coefficients a_i are real and so $\overline{a}_i = a_i$. □

The above theorem ensures that, with field $\mathbb{F} = \mathbb{C}$, every nonconstant polynomial of degree k has a list of exactly k roots $\{r_1, ..., r_k\}$. Some values may be repeated on this list, so the number of *distinct* roots is at least 1 and at most k. The number of times a distinct root is repeated in the list $\{r_1, ..., r_k\}$ is called the *algebraic multiplicity* of that root. The sum of the algebraic multiplicity of each distinct root is equal to the degree of the polynomial. If a polynomial has real coefficients and odd degree, part (c) implies that, since all roots come in complex conjugate pairs, there must be at least one real root r (since $\overline{r} = r$). This is consistent with the limit derivation of this same fact in the previous section.

13.3. Complex eigenvalues

Let A be a $n \times n$ matrix with entries in $\mathbb{C}$. Since every real number can be viewed as a complex number, matrices with all real entries can also be viewed as matrices with entries in $\mathbb{C}$. A scalar $\lambda \in \mathbb{C}$ is an eigenvalue of A if and only if

$$\det(A - \lambda I) = 0$$

Using (11.16), it can be shown that $\det(A - \lambda I)$ is a degree-n polynomial in λ of the form

$$\det(A - \lambda I) = (-1)^n \lambda^n + \theta_{n-1}\lambda^{n-1} + \cdots + \theta_1 \lambda + \theta_0 \quad \forall \lambda \in \mathbb{C}$$

for some complex-valued coefficients $\theta_{n-1}, ..., \theta_0$. The fundamental theorem of algebra ensures this polynomial has at least one (possibly complex-valued) root λ_1 (Theorem 13.1(a)). Therefore, *every square matrix with complex entries has at least one (possibly complex-valued) eigenvalue.* This proves Theorem 12.2(a).

Theorem 13.1(b) ensures the degree n polynomial $\det(A - \lambda I)$ can be factored

$$\det(A - \lambda I) = (-1)^n (\lambda - r_1)(\lambda - r_2)\cdots(\lambda - r_n) \quad \forall \lambda \in \mathbb{C}$$

where $\{r_1, ..., r_n\}$ is the list of exactly n roots, with some roots being repeated according to algebraic multiplicity. Recall that Theorem 12.2 ensures that A is similar to an upper triangular matrix U. It can be shown that the eigenvalues of A are on the n diagonal entries of U, these diagonal entries are precisely the roots $\{r_1, ..., r_n\}$ of the polynomial $\det(A - \lambda I)$, repetitions of each distinct eigenvalue r on the diagonal of U is exactly its algebraic multiplicity in the list of roots $\{r_1, ..., r_n\}$.

13.3.1. Examples

1. Consider
$$A = \begin{bmatrix} 1 & 2 & 5 \\ 0 & 1 & -2 \\ 0 & 0 & 4 \end{bmatrix} \quad , \quad \det(A - \lambda I) = (1-\lambda)^2(4-\lambda)$$
Eigenvalues are $\lambda_1 = 1, \lambda_2 = 1, \lambda_3 = 4$, so the eigenvalue 1 has algebraic multiplicity 2.
2. Consider
$$A = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \quad , \quad \det(A - \lambda I) = \lambda^2 + 1 = (\lambda - i)(\lambda + i) \tag{13.2}$$
Given a vector $x \in \mathbb{R}^2$, the matrix Ax rotates the vector 90 degrees clockwise. Hence, the real matrix A has no eigenvectors or eigenvalues if we restrict attention only to the real numbers. Indeed, the polynomial $\lambda^2 + 1$ has no roots $\lambda \in \mathbb{R}$. However, there are two complex roots: $\lambda_1 = i, \lambda_2 = -i$. The corresponding eigenvectors are any nonzero vectors in $Null(A - \lambda I)$ are are found by Gaussian elimination (using $\lambda \in \{\lambda_1, \lambda_2\}$):
$$A - \lambda I = \begin{bmatrix} -\lambda & 1 \\ -1 & -\lambda \end{bmatrix} \overset{(a)}{\to} \begin{bmatrix} -1 & -\lambda \\ -\lambda & 1 \end{bmatrix} \overset{(b)}{\to} \begin{bmatrix} -1 & -\lambda \\ 0 & 0 \end{bmatrix}$$
where step (a) swaps rows; step (b) uses the fact that we already know the rows are multiples of each other because λ makes the matrix $A - \lambda I$ not invertible. We look for an eigenvector $v = (c, 1)$. Then
$$-c - \lambda = 0 \implies c = -\lambda$$
so we obtain eigenvectors v_1, v_2 by
$$\lambda_1 = i \implies v_1 = \begin{bmatrix} -\lambda_1 \\ 1 \end{bmatrix} = \begin{bmatrix} -i \\ 1 \end{bmatrix} \quad , \quad \lambda_2 = -i \implies v_2 = \begin{bmatrix} -\lambda_2 \\ 1 \end{bmatrix} = \begin{bmatrix} i \\ 1 \end{bmatrix}$$
One can verify these are indeed eigenvectors:
$$\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} -i \\ 1 \end{bmatrix} = i \begin{bmatrix} -i \\ 1 \end{bmatrix} \quad , \quad \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} i \\ 1 \end{bmatrix} = -i \begin{bmatrix} i \\ 1 \end{bmatrix}$$

It follows that the matrix A is diagonalizable:

$$A = S\Lambda S^{-1} = \begin{bmatrix} -i & i \\ 1 & 1 \end{bmatrix}\begin{bmatrix} i & 0 \\ 0 & -i \end{bmatrix}\begin{bmatrix} -i & i \\ 1 & 1 \end{bmatrix}^{-1} \quad , \quad \begin{bmatrix} -i & i \\ 1 & 1 \end{bmatrix}^{-1} = \frac{1}{2}\begin{bmatrix} i & 1 \\ -i & 1 \end{bmatrix}$$

13.4. Eigenvalues for linear operators over field $\mathbb{C}$

Recall that if V is a vector space over a field, a scalar λ is an eigenvalue of a linear operator $T : V \to V$ if there is a nonzero vector $v \in V$ such that $T(v) = \lambda v$ (the nonzero vector v is called an eigenvector, and (λ, v) is an eigenpair).

LEMMA 13.1. *Let V be a vector space over field $\mathbb{C}$. Assume $\dim(V)$ is a positive integer. Let $T : V \to V$ be a linear transformation. Then T has a (possibly complex-valued) eigenvalue.*

PROOF. Let $n = \dim(V)$ and let $\{v_1, ..., v_n\}$ be a basis for V. Let $A = Mat(T)$ be the matrix of T with respect to this basis. So $A = (a_{ij})$ is a $n \times n$ matrix with complex entries a_{ij} that satisfy

$$T(v_j) = \sum_{i=1}^{n} a_{ij} v_i \quad \forall j \in \{1, \ldots, n\} \tag{13.3}$$

We already know A has an eigenvalue $\lambda \in \mathbb{C}$. Let $x = (x_1, ..., x_n) \in \mathbb{C}^n$ be the corresponding eigenvector. Since x is an eigenvector we know $x \neq 0$. The equation $Ax = \lambda x$ implies

$$\sum_{j=1}^{n} a_{ij} x_j = \lambda x_i \quad \forall i \in \{1, \ldots, n\} \tag{13.4}$$

We now show λ is also an eigenvalue of T. Define the following vector $v \in V$:

$$v = \sum_{j=1}^{n} x_j v_j \tag{13.5}$$

Note that $v \neq 0$. (Indeed, if $v = 0$ then linear independence of $\{v_1, ..., v_n\}$ would imply $x_j = 0$ for all j, which is impossible since $x \neq 0$.) Then

$$\begin{aligned} T(v) &= T\left(\sum_{j=1}^{n} x_j v_j\right) \\ &= \sum_{j=1}^{n} x_j T(v_j) \\ &\overset{(a)}{=} \sum_{j=1}^{n} x_j \left[\sum_{i=1}^{n} a_{ij} v_i\right] \end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} v_i \left[\sum_{j=1}^{n} a_{ij} x_j \right] \\
&\overset{(b)}{=} \sum_{i=1}^{n} v_i \left[\lambda x_i \right] \\
&\overset{(c)}{=} \lambda v
\end{aligned}
$$

where equality (a) holds by (13.3); equality (b) holds by (13.4); equality (c) holds by definition of v in (13.5). Thus, v is an eigenvector of T with eigenvalue λ. □

13.5. Invariant subspaces

Let V be a complex vector space. Let $T : V \to V$ be a linear transformation. A subspace $S \subseteq V$ is said to be *invariant under* T if

$$v \in S \implies T(v) \in S$$

An invariant subspace is said to be nontrivial if it contains at least one nonzero vector.

THEOREM 13.2. *Let V be a complex vector space. Let $T : V \to V$ be a linear transformation. If S is a nontrivial and finite-dimensional subspace of V that is invariant under T, then T has an eigenvector in S. In particular, if A is a $n \times n$ matrix with complex entries and $S \subseteq \mathbb{C}^n$ is a nontrivial subspace that satisfies*

$$x \in S \implies Ax \in S$$

then A has an eigenvector in S.

PROOF. Note that S is itself a complex vector space with $\dim(S)$ being a positive integer. Define the linear transformation $\tilde{T} : S \to S$ by

$$\tilde{T}(x) = T(x) \quad \forall x \in S$$

This indeed maps every vector $x \in S$ to another vector $T(x) \in S$ because S is invariant under T. By Lemma 13.1, we know $\tilde{T}$ has an eigenvalue $\lambda \in \mathbb{C}$. So there is a nonzero vector $v \in S$ such that $\tilde{T}(v) = \lambda v$. Thus, $T(v) = \lambda v$, so v is also an eigenvector of T. □

As an example consider the real and symmetric matrix

$$A = \begin{bmatrix} 1 & 3 & -7 \\ 3 & 0 & 0 \\ -7 & 0 & 2 \end{bmatrix} \tag{13.6}$$

Observe that $A^\top = A$, so A is symmetric. It can be shown that $\lambda_1 = 9$ is an eigenvalue of A. Let $v_1 \in \mathbb{C}^3$ be the corresponding eigenvector. Define

$$S = Span\{v_1\}^\perp$$

That is, S is the 2-dimensional subspace of all vectors orthogonal to v_1, where we use the standard inner product $\langle x, y\rangle = x^H y$ for all $x, y \in \mathbb{C}^3$. It holds that S is an invariant subspace because if $x \in S$ then

$$\langle x, v_1\rangle = x^H v_1 = 0$$

and so

$$\langle Ax, v_1\rangle = (Ax)^H v_1 = x^H A^H v_1 \overset{(a)}{=} x^H A v_1 = x^H(\lambda v_1) = \lambda(x^H v_1) = 0$$

where (a) holds because $A^H = A$ because A is real and symmetric. Thus, if $x \in S$ then $Ax \in S$ (so S is indeed invariant). It follows that A has another eigenvector in S. In particular, A has an eigenvector v_2 that is orthogonal to our first eigenvector v_1. One can then define the 1-dimensional subspace

$$U = Span\{v_1, v_2\}^\perp$$

and a similar argument shows that U is invariant. So A has an eigenvector $v_3 \in U$. In particular, A has an *orthogonal basis of eigenvectors* $\{v_1, v_2, v_3\}$.

We can always normalize the eigenvectors:

$$\left\{\frac{v_1}{||v_1||}, \frac{v_2}{||v_2||}, \frac{v_3}{||v_3||}\right\}$$

so we can say that A has an *orthonormal basis of eigenvectors*. This fact about an orthonormal basis of eigenvectors generalizes to any symmetric matrix A with real entries (see Chapter 15). It also holds for *skew symmetric matrices* such as the example matrix in (13.2).

13.6. Polynomial functions of matrices

Let A be a $n \times n$ matrix with complex entries. Suppose (λ, v) is an eigenpair of A. Questions:

1. What is an eigenpair of A^2?
2. What is an eigenpair of A^3?
3. What is an eigenpair of $5A^3 - 3A^2 + A + I$?
4. If A is invertible, what is an eigenpair of A^{-1}?
5. If A is invertible, what is an eigenpair of $7A + 3A^{-1} + 2A^{-2}$? (where $A^{-2} = (A^{-1})^2$)

It is easy to see that if $Av = \lambda v$, then multiplying both sides by A gives $A^2 v = \lambda A v = \lambda^2 v$.

This answers the first question: If (λ, v) is an eigenpair of A, then (λ^2, v) is an eigenpair of A^2. In fact, this holds when we use any field $\mathbb{F}$, although let us continue to assume $\mathbb{F} = \mathbb{C}$ here. The answers to questions 2 and 3 are similar: The eigenvector v remains an eigenvector, and the eigenvalue λ changes to λ^3 and $5\lambda^3 - 3\lambda^2 + \lambda + 1$, respectively. The fourth question is somewhat different: We start with $Av = \lambda v$; since we assume A^{-1} exists for this fourth question we can multiply both sides by A^{-1} to obtain $v = \lambda A^{-1}v$; we also know $\lambda \neq 0$ (why?) and so we divide both sides by λ to conclude $A^{-1}v = \lambda^{-1}v$. So the new eigenpair is (λ^{-1}, v). Then the fifth question has answer $(7\lambda + 3\lambda^{-1} + 2\lambda^{-2}, v)$. There is a clear pattern here.

Fix k as a positive integer and let $p : \mathbb{C} \to \mathbb{C}$ be a degree-k polynomial:

$$p(z) = c_k z^k + c_{k-1} z^{k-1} + \cdots + c_1 z + c_0 \quad \forall z \in \mathbb{C} \tag{13.7}$$

(where we assume $c_k \neq 0$). Now let A be any $n \times n$ matrix with complex entries. We define $p(A)$ by

$$p(A) = c_k A^k + c_{k-1} A^{k-1} + \cdots + c_1 A + c_0 I$$

where I is the $n \times n$ identity matrix. The above discussion suggests that if λ is an eigenvalue of A, then $p(\lambda)$ is an eigenvalue of $p(A)$. (This is true and is proven in Theorem 13.3 given below.) However, it is not obvious if these are the *only* eigenvalues of $p(A)$. It turns out that this is also true, and this is called the *spectral mapping theorem* (formally stated and proven below).

An intriguing thing we can do with polynomials applied to matrices is *factor them* according to the standard polynomial factorization. We know that all degree k polynomials can be factored (over the field $\mathbb{C}$) into k factors with roots $z_1, \ldots, z_k$ (repeated according to algebraic multiplicity). So the polynomial $p(z)$ in (13.7) can be factored:

$$p(z) = c_k(z - z_1)(z - z_2)\cdots(z - z_k) \quad \forall z \in \mathbb{C}$$

It follows that for any square matrix A with complex entries, $p(A)$ can be expressed in this factored form:

$$\begin{aligned} p(A) &= c_k A^k + c_{k-1} A^{k-1} + \cdots + c_1 A + c_0 I \\ &= c_k(A - z_1 I)(A - z_2 I)\cdots(A - z_{k-1} I)(A - z_k I) \end{aligned}$$

Further, since the factorization of $p(z)$ does not depend on the ordering of the factors, the factorization of $p(A)$ also does not depend on this ordering! For example, we can reverse the ordering of roots:

$$p(A) = c_k(A - z_k I)(A - z_{k-1} I)\cdots(A - z_2 I)(A - z_1 I)$$

This leads to the following lemma.

LEMMA 13.2. *Let $p : \mathbb{C} \to \mathbb{C}$ be a polynomial function of degree $k \geq 1$. Let A be a $n \times n$ matrix with complex entries. Let w be* any *nonzero vector in $\mathbb{C}^n$. If $p(A)w = 0$ then at least one root of the polynomial $p(z)$ is an eigenvalue of A.*

PROOF. Suppose w is a nonzero vector in $\mathbb{C}^n$ such that

$$p(A)w = 0$$

Since $k \geq 1$, the degree-k polynomial has some factorization

$$p(z) = c_k(z - z_k)(z - z_{k-1})\cdots(z - z_1)$$

for some $c_k, z_1, \ldots, z_k \in \mathbb{C}$ (with $c_k \neq 0$). From this we obtain

$$0 = p(A)w = c_k(A - z_kI)(A - z_{k-1}I)\cdots(A - z_2I)(A - z_1I)w$$

which means that one of the following k possibilities is true:

- $(A - z_1I)w = 0$ so w is an eigenvector of A with eigenvalue z_1.
- $(A - z_1I)w \neq 0$ but $(A - z_2I)(A - z_1I)w = 0$. So $(A - z_1I)w$ is an eigenvector of A with eigenvalue z_2.
- $(A - z_2I)(A - z_1I)w \neq 0$ but $(A - z_3I)(A - z_2I)(A - z_1I)w = 0$. So $(A - z_2I)(A - z_1I)w$ is an eigenvector of A with eigenvalue z_3.

  ⋮
- $(A - z_{k-1}I)(A - z_{k-2}I)\cdots(A - z_1I)w$ is an eigenvector of A with eigenvalue z_k.

and so at least one root of $p(z)$ is an eigenvalue of A. □

Lemma 13.2 can be used as an elegant argument that every square complex matrix A has a (possibly complex) eigenvalue, and the argument does not require the determinant concept (see also Axler (2024)): Take *any* nonzero $w \in \mathbb{C}^n$. Consider the $n + 1$ vectors

$$\{w, Aw, A^2w, A^3, \ldots, A^nw\}$$

These are $n + 1$ vectors in the n-dimensional vector space $\mathbb{C}^n$, so they cannot be linearly independent. There must be complex coefficients c_i (not all zero) such that

$$\sum_{i=0}^{n} c_iA^iw = 0 \tag{13.8}$$

where A^0 is defined as I. Let $k \in \{1, \ldots, n\}$ be the largest index i for which $c_i \neq 0$ (we cannot

have $k = 0$ since then $c_0 \neq 0$ and (13.8) would reduce to $c_0 w = 0$, which is impossible since $w \neq 0$). Define the degree-k polynomial

$$p(z) = c_0 + c_1 z + c_2 z^2 + \ldots + c_k z^k$$

Then (13.8) becomes $p(A)w = 0$, which (by Lemma 13.2) means at least one of the roots of the degree-k polynomial $p(z)$ is an eigenvalue of A. In particular, A has an eigenvalue!

Lemma 13.2 can also be used to prove the following spectral mapping theorem.

THEOREM 13.3. *(Spectral mapping) Fix n, k as positive integers. Let $p : \mathbb{C} \to \mathbb{C}$ be a degree-k polynomial and let A be a $n \times n$ matrix with complex entries. Here we consider the field $\mathbb{C}$ (so eigenvalues can be complex numbers). Then*

a) If (λ, v) is an eigenpair of A, then $(p(\lambda), v)$ is an eigenpair of $p(A)$.

b) If μ is an eigenvalue of $p(A)$, then there is a $\lambda \in \mathbb{C}$ such that λ is an eigenvalue of A and $p(\lambda) = \mu$. Thus, the set of all eigenvalues of $p(A)$ is given by:

$$\Lambda(p(A)) = \{p(\lambda) : \lambda \text{ is an eigenvalue of } A\}$$

PROOF. To prove part (a), suppose (λ, v) is an eigenpair of A. We have

$$p(z) = c_k z^k + c_{k-1} z^{k-1} + \cdots + c_1 z + c_0 \quad \forall z \in \mathbb{C}$$

for some $c_0, c_1, \ldots, c_k \in \mathbb{C}$, so

$$p(A) = c_k A^k + c_{k-1} A^{k-1} + \cdots + c_1 A + c_0 I$$

Since $Av = \lambda v$, we know $A^j v = \lambda^j v$ for all $j \in \{1, 2, 3, \ldots\}$. Therefore

$$\begin{aligned} p(A)v &= c_k A^k v + c_{k-1} A^{k-1} v + \cdots + c_1 Av + c_0 v \\ &= c_k \lambda^k v + c_{k-1} \lambda^{k-1} v + \cdots + c_1 \lambda v + c_0 v \\ &= p(\lambda) v \end{aligned}$$

So $(p(\lambda), v)$ is an eigenpair of $p(A)$.

To prove part (b), suppose μ is an eigenvalue of $p(A)$. Let v be the corresponding eigenvector. So $p(A)v = \mu v$ and

$$(p(A) - \mu I)v = 0$$

Define the degree $k \geq 1$ polynomial function $q(z) = p(z) - \mu$ for $z \in \mathbb{C}$. The above equality means

$$q(A)v = 0$$

so Lemma 13.2 ensures at least one root of $q(z)$ is an eigenvalue of A. Let $\lambda \in \mathbb{C}$ be a root

of $q(z)$ that is an eigenvalue of A. Then $0 = q(\lambda) = p(\lambda) - \mu$. That is, λ is an eigenvalue of A and $p(\lambda) = \mu$. This proves the first sentence of part (b). The second sentence follows immediately from this and part (a). □

While every eigenvector of A is also an eigenvector of $p(A)$, the reverse is not necessarily true. For example, consider $p(z) = z^2$ for $z \in \mathbb{C}$ and

$$A = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix} \quad , \quad p(A) = A^2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Every nonzero vector in $\mathbb{C}^2$ is an eigenvector of $p(A)$, but the same is not true of A.

13.7. Exercises

EXERCISE 13.1. *Consider the real matrix A:*

$$A = \begin{bmatrix} 2 & -2 \\ 4 & -1 \end{bmatrix}$$

a) Find λ_1, λ_2.

b) Find $\sum_{i=1}^{2} \lambda_i$ *and* $\prod_{i=1}^{2} \lambda_i$

EXERCISE 13.2. *Consider the real matrix A:*

$$A = \begin{bmatrix} -7 & 1 \\ 4 & 4 \end{bmatrix}$$

a) Find λ_1, λ_2.

c) Find $\sum_{i=1}^{2} \lambda_i$ *and* $\prod_{i=1}^{2} \lambda_i$

EXERCISE 13.3. *Consider the real matrix A:*

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

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to obtain $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

EXERCISE 13.4. *Consider the real matrix A:*

$$A = \begin{bmatrix} 2 & 2 & 0 \\ 4 & 10 & 3 \\ -5 & 2 & 1 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to obtain $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

EXERCISE 13.5. *(Repeating a row) Consider the real matrix A:*

$$A = \begin{bmatrix} 18 & -6 \\ 18 & -6 \end{bmatrix}$$

a)Find λ_1, λ_2.

b) Find $\sum_{i=1}^{2} \lambda_i$ *and* $\prod_{i=1}^{2} \lambda_i$.

c) Diagonalize A.

EXERCISE 13.6. *(Repeating a column) Consider the real matrix A:*

$$A = \begin{bmatrix} 1 & 1 \\ 4 & 4 \end{bmatrix}$$

a)Find λ_1, λ_2.

b) Find $\sum_{i=1}^{2} \lambda_i$ *and* $\prod_{i=1}^{2} \lambda_i$.

c) Diagonalize A.

EXERCISE 13.7. *(Repeating a row) Consider the real matrix A:*

$$A = \begin{bmatrix} 4 & 0 & 5 \\ 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

EXERCISE 13.8. *(Repeating a column) Consider the real matrix A:*

$$A = \begin{bmatrix} 4 & 9 & 9 \\ 1 & 9 & 9 \\ 1 & 9 & 9 \end{bmatrix}$$

a) Find the degree 3 polynomial $f(\lambda) = \det(A - \lambda I)$.

b) Numerically solve to find $\lambda_1, \lambda_2, \lambda_3$.

c) Find $\sum_{i=1}^{3} \lambda_i$ *and* $\prod_{i=1}^{3} \lambda_i$

EXERCISE 13.9. *Let A be a square matrix with entries in some field* $\mathbb{F}$. *Let* (λ, v) *be an eigenpair of A.*

a) Find an eigenpair of $-A$.

b) Find an eigenpair of A^2.

c) Fix $c \in \mathbb{F}$. *Find an eigenpair of* $A^3 + A^2 + cA$.

d) Let S be an invertible matrix of the same size as A. Find an eigenpair of SAS^{-1}.

e) Assume A is invertible. Find an eigenpair of A^{-1}.

EXERCISE 13.10. *Let V be a vector space over a field* $\mathbb{F}$. *Let* $T : V \to V$ *be a linear transformation. Suppose that* (λ, v) *is an eigenpair of T, so that* $\lambda \in \mathbb{F}$, $v \in V$, $v \neq 0$, *and* $T(v) = \lambda v$.

a) Define $G = T^2$. *That is,* $G : V \to V$ *is given by* $G(x) = T(T(x))$ *for all* $x \in V$. *Find an eigenpair of G.*

b) Fix $c_0, c_1, c_2 \in \mathbb{F}$. *Define* $H = c_0 I + c_1 T + c_2 T^2$, *so* $H(x) = c_0 x + c_1 T(x) + c_2 T^2(x)$ *for all* $x \in V$. *Find an eigenpair of H.*

EXERCISE 13.11. *Suppose A is a* $n \times n$ *matrix with entries in some field* $\mathbb{F}$. *Fix* $c_1, c_2, c_3 \in \mathbb{F}$. *Fix* $v \in \mathbb{F}^n$ *such that* $v \neq 0$. *Suppose that*

$$(A - c_1 I)(A - c_2 I)(A - c_3 I)v = 0$$

a) Show that A must have an eigenvalue $\lambda \in \{c_1, c_2, c_3\}$.

b) Give an example where c_3 *is not an eigenvalue of A, but* c_2 *is.*

EXERCISE 13.12. *Let A be a* $n \times n$ *matrix with entries in some field* $\mathbb{F}$. *Fix* $c_0, c_1 \in \mathbb{F}$. *Consider the quadratic polynomial* $p : \mathbb{F} \to \mathbb{F}$ *defined by*

$$p(x) = x^2 + c_1 x + c_0 \quad \forall x \in \mathbb{F}$$

Define

$$p(A) = A^2 + c_1 A + c_0 I$$

where I is the $n \times n$ *identity matrix.*

a) Suppose (λ, v) is an eigenpair of A. Find an eigenpair of $p(A)$.

b) Suppose the polynomial can be factored, so that

$$p(x) = (x - z_1)(x - z_2) \quad \forall x \in \mathbb{F}$$

for some roots $z_1, z_2 \in \mathbb{F}$. Prove that

$$p(A) = (A - z_1 I)(A - z_2 I)$$

c) Assume the conditions of part (b). Let v be a nonzero vector such that $p(A)v = 0$. Show that either z_1 or z_2 must be an eigenvalue of A.

EXERCISE 13.13. *(Consequences of factoring) Let n, k be positive integers. Let A be a $n \times n$ matrix with entries in some field $\mathbb{F}$. Let $p : \mathbb{F} \to \mathbb{F}$ be a degree-k polynomial:*

$$p(x) = x^k + c_{k-1}x^{k-1} + c_{k-2}x^{k-2} + \cdots + c_1 x + c_0 \quad \forall x \in \mathbb{F}$$

Define

$$p(A) = A^k + c_{k-1}A^{k-1} + c_{k-2}A^{k-2} + \cdots + c_1 A + c_0 I$$

where I is the $n \times n$ identity matrix. Suppose the polynomial p can be factored:

$$p(x) = (x - z_1)(x - z_2)\cdots(x - z_k) \quad \forall x \in \mathbb{F}$$

for some roots $z_1, \ldots, z_k \in \mathbb{F}$.

a) Use induction to prove that $p(A) = (A - z_1 I)(A - z_2 I)\cdots(A - z_k I)$.

b) Suppose there is a nonzero vector $v \in \mathbb{F}^n$ such that $p(A)v = 0$. Prove that A has an eigenvalue $\lambda \in \{z_1, \ldots, z_k\}$.

c) Fix any nonzero vector $v \in \mathbb{F}^n$ and consider the sequence of vectors $\{v, Av, A^2v, \ldots, A^k v\}$. Show there must be an integer $k \in \{2, \ldots, n\}$ such that $A^k v \in Span\{v, Av, \ldots, A^{k-1}v\}$.

d) Using the setup of part (c), explain why there are scalar coefficients d_i such that

$$A^k v = d_0 v + d_1 Av + d_2 A^2 v + \ldots + d_{k-1}A^{k-1}v$$

and so $q(A)v = 0$, where q is the degree-k polynomial

$$q(x) = x^k - (d_0 + d_1 x + d_2 x^2 + \ldots + d_{k-1}x^{k-1}) \quad \forall x \in \mathbb{F}$$

Explain why, if q can be factored into a product of k roots, then at least one of those roots must be an eigenvalue of A. It can be shown that if $\mathbb{F} = \mathbb{C}$ then every degree k polynomial can be factored into a product of k (possibly repeated) roots. Therefore, in the complex field, every square matrix has at least one eigenvalue, and such an eigenvalue can be found by starting with any nonzero

vector $v \in \mathbb{C}^n$, considering $\{v, Av, A^2v, ..., A^{k-1}v\}$, and factoring the corresponding polynomial as described in previous parts of this problem.

EXERCISE 13.14. *(Commuting operators over the complex field share an eigenvector) Fix $n \in \mathbb{N}$. Let V be a n-dimensional vector space over field $\mathbb{C}$ and let $T : V \to V$ and $G : V \to V$ be linear operators. Suppose that T and G commute, so*

$$T(G(x)) = G(T(x)) \quad \forall x \in V$$

We want to show there is a vector $w \in V$ such that w is an eigenvector for both T and G (with possibly different eigenvalues). Let (λ, v) be an eigenpair of T (we know such an eigenpair exists because this is a complex vector space of finite dimension $n \geq 1$). Define $S = Null(T - \lambda I)$, where $I : V \to V$ is the identity operator $I(x) = x$ for all $x \in V$.

a) Explain why S is nontrivial, and why all nonzero vectors in S are eigenvectors of T.

b) Show that S is an invariant subspace under G (meaning that $x \in S \implies G(x) \in S$).

c) Use parts (a) and (b) to conclude S contains a vector that is an eigenvector for both T and G.

EXERCISE 13.15. *(Eigenvectors for linear operators) Fix $n \in \mathbb{N}$. Let V be a n-dimensional vector space over field $\mathbb{F}$. Let $T : V \to V$ be a linear transformation. Suppose that T has eigenvectors $\{v_1, ..., v_n\}$ that form a basis for V, with corresponding (possibly repeated) eigenvalues $\lambda_1, ..., \lambda_n$. Suppose x is some other eigenvector of T with eigenvalue $\mu \in \mathbb{F}$.*

a) Explain why $\mu \in \{\lambda_1, ..., \lambda_n\}$.

b) Show that we can form a basis for V of the form $\{x, w_2, ..., w_n\}$ for some eigenvectors w_i for $i \in \{2, ..., n\}$.

CHAPTER 14.

The ODE $x'(t) = Ax(t)$

Eigenvalues and eigenvectors are useful for solving the following vector-based ordinary differential equation (ODE):

$$x'(t) = Ax(t) \quad \forall t \in \mathbb{R} \tag{14.1}$$

where A is a given $n \times n$ matrix and $x(t) = (x_1(t), \ldots, x_n(t))$. Note that $x(t)$ is treated as a column vector for the purposes of matrix multiplication in (14.1). The derivative $x'(t)$ is taken componentwise:

$$x'(t) = (x_1'(t), \ldots, x_n'(t)) = \left(\frac{d}{dt}x_1(t), \ldots, \frac{d}{dt}x_n(t)\right)$$

assuming that each component function $x_i(t)$ is differentiable. A function $x(t)$ that satisfies the ODE (14.1) for all $t \in \mathbb{R}$ is called a *solution* to the ODE. It is clear that the always-zero function ($x(t) = 0$ for all t) is a solution. The always-zero function has initial condition $x(0) = 0$. Typically we are given a nonzero initial condition $x(0) = x_0$, where x_0 is a given vector in $\mathbb{R}^n$, and we want to find a solution that matches that initial condition.

For example, suppose we want to find a function $x(t) = (x_1(t), x_2(t))$ that takes values in $\mathbb{R}^2$ and satisfies

$$\begin{aligned} x_1'(t) &= -3x_1(t) - 2x_2(t) \quad \forall t \in \mathbb{R} \\ x_2'(t) &= x_1(t) - x_2(t) \quad \forall t \in \mathbb{R} \end{aligned}$$

with initial condition $(x_1(0), x_2(0)) = (1, 1)$. This system has the form $x'(t) = Ax(t)$, with $x(0) = x_0$, for

$$A = \begin{bmatrix} -3 & -2 \\ 1 & -1 \end{bmatrix} \quad , \quad x_0 = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

The solution for this specific example is presented in Section 14.6.

14.1. The subspace of solutions

We are primarily interested in the case when A has real entries and $x(t)$ has real components. However, it is useful to allow the $n \times n$ matrix A to have complex components, and to allow solutions $x(t)$ take values in $\mathbb{C}^n$ for all t. A function $x(t) \in \mathbb{C}^n$ has real and imaginary parts, and its derivative is defined by the derivatives of the real and imaginary parts:

$$x(t) = w(t) + iz(t)$$
$$x'(t) = w'(t) + iz'(t)$$

where $w(t)$ and $z(t)$ take values in $\mathbb{R}^n$ for all t. The function $x(t)$ is differentiable if and only if its real and imaginary parts $w(t), z(t)$ are both differentiable. If A is a real $n \times n$ matrix, it can be shown that a complex valued function satisfies the ODE if and only if both its real and imaginary parts satisfy the ODE (Exercise 14.8).

Let V be the complex vector space of all differentiable functions $x(t)$ that take values in $\mathbb{C}^n$ over time $t \in \mathbb{R}$. We search for solutions to the ODE (14.1) in the space V. The set of all solutions to the ODE forms a subspace of V. Specifically, if $x(t)$ and $y(t)$ are two functions that satisfy (14.1), then any linear combination of $x(t)$ and $y(t)$ also satisfies (14.1) (see Exercise 14.13).

14.2. Existence and uniqueness

Given initial condition $x_0 \in \mathbb{C}^n$, the goal is to find a function $x(t)$ that satisfies the ODE (14.1) and also satisfies $x(0) = x_0$. Because the ODE (14.1) has a linear structure, a fundamental *existence and uniqueness theorem* ensures that for every $x_0 \in \mathbb{C}^n$, there exists one and only one function $x(t)$ that satisfies (14.1) and has $x(0) = x_0$. Further, if the initial condition has all real components, then the resulting solution has all real components for all time. That is,

$$x_0 \in \mathbb{R}^n \implies x(t) \in \mathbb{R}^n \quad \forall t \in \mathbb{R}$$

14.3. Solutions from eigenpairs of A

Here is the key to finding solutions to (14.1): If (λ, v) is an eigenpair of A (where $\lambda \in \mathbb{C}$ and $v \in \mathbb{C}^n$) then the following function $x(t)$ is a particular solution to the ODE (14.1):

$$x(t) = ve^{\lambda t} \quad \forall t \in \mathbb{R}$$

where

$$v = (v_1, v_2, ..., v_n)$$
$$ve^{\lambda t} = (v_1 e^{\lambda t}, v_2 e^{\lambda t}, ..., v_n e^{\lambda t})$$

This can easily be seen as follows:

$$\begin{aligned} x(t) = ve^{\lambda t} \implies x'(t) &= \frac{d}{dt}(v_1 e^{\lambda t}, v_2 e^{\lambda t} ..., v_n e^{\lambda t}) \\ &= (v_1 \lambda e^{\lambda t}, v_2 \lambda e^{\lambda t}, ..., v_n \lambda e^{\lambda t}) \\ &= \lambda v e^{\lambda t} \\ &\overset{(a)}{=} (Av)e^{\lambda t} \\ &\overset{(b)}{=} Ax(t) \end{aligned}$$

where equality (a) holds because $Av = \lambda v$; equality (b) holds because we have defined $x(t) = ve^{\lambda t}$.

Since $ve^{\lambda t}$ is a solution to the ODE (14.1), any scalar multiple of this is also a solution. This provides an infinite number of new solutions. However, all of these new solutions have initial conditions that are scalar multiples of v. If our desired initial condition x_0 is *not* a scalar multiple of v, then the single solution $ve^{\lambda t}$ is not good enough: We need to find n different solutions that have n linearly independent initial conditions $\{v_1, ..., v_n\}$. Then we can take linear combinations of these solutions to fit any desired initial condition $x_0 \in \mathbb{C}^n$.

14.4. The case when A is diagonalizable

The easiest situation is when the $n \times n$ matrix A is *diagonalizable*

$$A = S\Lambda S^{-1}$$

where the n columns of S consist of n linearly independent eigenvectors $v_1, ..., v_n$ (which are a basis for $\mathbb{C}^n$) with corresponding eigenvalues $\lambda_1, ..., \lambda_n$ (scalars in $\mathbb{C}$) that form the

entries of the diagonal matrix Λ. This gives n distinct eigenpairs (λ_i, v_i) of the matrix A. Therefore, for each $i \in \{1, \ldots, n\}$ we have the following solution to the ODE (14.1):

$$x_i(t) = v_i e^{\lambda_i t} \quad \forall t$$

where $x_i(t) \in \mathbb{C}^n$ now represents the ith vector-valued solution (not the ith component of a vector-valued function). If λ_i and v_i are both real, then $x_i(t)$ has real components for all time t. Given a desired initial condition $x_0 \in \mathbb{C}^n$, we can take linear combinations of $x_1(t), \ldots, x_n(t)$ to find the unique solution to the ODE with that initial condition. Specifically, define a new solution $x(t)$ by

$$\boxed{x(t) = \sum_{i=1}^{n} \theta_i v_i e^{\lambda_i t} \quad \forall t \in \mathbb{R}} \tag{14.2}$$

where $\theta_1, \ldots, \theta_n$ are scalar weights (generally complex-valued). We have $x(0) = x_0$ if and only if

$$x(0) = \sum_{i=1}^{n} \theta_i v_i = x_0$$

Therefore, we choose the scalars θ_i to solve

$$S\theta = x_0$$

where S is the matrix with columns given by the eigenvectors v_i, and $\theta = (\theta_1, \ldots, \theta_n)$. It follows that

$$\boxed{\theta = S^{-1} x_0}$$

Using these scalars θ, the function $x(t)$ defined in (14.2) satisfies the ODE (14.1) and matches the desired initial condition $x(0) = x_0$. Therefore, the existence and uniqueness theorem ensures $x(t)$ is the unique solution with that initial condition. This unique solution (14.2) can be equivalently written in matrix form

$$\boxed{x(t) = S e^{\Lambda t} S^{-1} x_0 \quad \forall t \in \mathbb{R}} \tag{14.3}$$

where, for the diagonal matrix Λ, we define

$$e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & 0 & \cdots & 0 \\ 0 & 0 & e^{\lambda_3 t} & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} \tag{14.4}$$

Recall that if $\lambda = a + bi$ then

$$e^{\lambda t} = e^{(a+bi)t} = e^{at}(\cos(bt) + i\sin(bt)) \quad \forall t \in \mathbb{R} \tag{14.5}$$

It follows that

$$(a < 0) \implies \lim_{t\to\infty} e^{\lambda t} = 0$$

Thus, in this diagonalizable case, if all eigenvalues of A have negative real part, then all terms in the matrix (14.4) converge to 0 as $t \to \infty$, so $x(t) \to 0$ (regardless of the initial condition $x(0) \in \mathbb{C}^n$). Even when A is *not* diagonalizable, if all eigenvalues of A have negative real part, it can be shown that $\lim_{t\to\infty} x(t) = 0$ for all initial conditions $x(0) \in \mathbb{C}^n$ (see Section 14.9). On the other hand, if at least one eigenvalue of A has a nonnegative real part, then there exists an initial condition $v \in \mathbb{C}^n$ for which $x(t)$ does *not* converge to zero as $t \to \infty$. [Indeed, given an eigenpair (λ, v) where λ has nonnegative real part, just consider the particular solution $x(t) = ve^{\lambda t}$ and use the fact (14.5).] For this reason, the ODE $x'(t) = Ax(t)$ is called *stable* if all eigenvalues of A have negative real part.

14.5. Example: Diagonalizable with real eigenvalues

Consider

$$A = \begin{bmatrix} 4 & -1 \\ -2 & 3 \end{bmatrix} \quad , \quad x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{14.6}$$

To solve $x'(t) = Ax(t)$ with $x(0) = x_0$, we find the eigenvalues:

$$\det(A - \lambda I) = 0 \iff (4 - \lambda)(3 - \lambda) - 2 = 0$$

and the eigenvalues are

$$\lambda_1 = 5$$
$$\lambda_2 = 2$$

The matrix has two distinct eigenvalues, so it is diagonalizable and has a basis of eigenvectors. Since A is real and the eigenvalues are real, the eigenvectors can be chosen to be real. To find an eigenvector for eigenvalue λ (where $\lambda = \lambda_1$ or $\lambda = \lambda_2$) we find a nontrivial solution to $(A - \lambda I)v = 0$. By Gaussian elimination we have

$$\begin{bmatrix} 4-\lambda & -1 \\ -2 & 3-\lambda \end{bmatrix} \to \begin{bmatrix} -2 & 3-\lambda \\ 4-\lambda & -1 \end{bmatrix} \to \begin{bmatrix} \boxed{-2} & 3-\lambda \\ 0 & 0 \end{bmatrix}$$

where the last elimination step follows because the rows must be multiples of each other (since we know $A - \lambda I$ is not invertible). The first elimination step of swapping the rows was not needed, but was for convenience of placing the nonzero constant -2 in the upper left corner to act as a pivot. We have a single free variable. For convenience, we search for an eigenvector of the form $(a, 2)$:

$$\begin{bmatrix} -2 & 3-\lambda \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies -2a + 2(3-\lambda) = 0 \implies a = 3 - \lambda$$

so the eigenvectors are $v_i = (3 - \lambda_i, 2)$ for $i \in \{1, 2\}$, that is

$$\lambda_1 = 5 \implies v_1 = \begin{bmatrix} 3-\lambda_1 \\ 2 \end{bmatrix} = \begin{bmatrix} -2 \\ 2 \end{bmatrix}$$

$$\lambda_2 = 2 \implies v_2 = \begin{bmatrix} 3-\lambda_2 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$$

Then a general solution to $x'(t) = Ax(t)$ is

$$x(t) = \theta_1 \begin{bmatrix} -2 \\ 2 \end{bmatrix} e^{5t} + \theta_2 \begin{bmatrix} 1 \\ 2 \end{bmatrix} e^{2t} \quad \forall t \in \mathbb{R}$$

where θ_1, θ_2 are any scalar weights. To match the initial condition $x(0) = (1, 1)$ we have

$$x(0) = \theta_1 \begin{bmatrix} -2 \\ 2 \end{bmatrix} + \theta_2 \begin{bmatrix} 1 \\ 2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

and so

$$\begin{bmatrix} -2 & 1 \\ 2 & 2 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \implies \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 2 & 2 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1/6 \\ 2/3 \end{bmatrix}$$

Therefore

$$\boxed{x(t) = (-1/6)\begin{bmatrix} -2 \\ 2 \end{bmatrix} e^{5t} + (2/3)\begin{bmatrix} 1 \\ 2 \end{bmatrix} e^{2t} \quad \forall t \in \mathbb{R}}$$

One can verify that $x(0) = (1, 1)$ and $x'(t) = Ax(t)$. Equivalent forms of the solution are

$$x(t) = \begin{bmatrix} (1/3)e^{5t} + (2/3)e^{2t} \\ (-1/3)e^{5t} + (4/3)e^{2t} \end{bmatrix} \quad \forall t \in \mathbb{R}$$

$$x(t) = \underbrace{\begin{bmatrix} -2 & 1 \\ 2 & 2 \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} e^{5t} & 0 \\ 0 & e^{2t} \end{bmatrix}}_{e^{\Lambda t}} \underbrace{\begin{bmatrix} -2 & 1 \\ 2 & 2 \end{bmatrix}^{-1}}_{S^{-1}} \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{x_0} \quad \forall t \in \mathbb{R}$$

The reader can verify that any alternative choice of eigenvectors v_1, v_2 for $\lambda_1 = 5, \lambda_2 = 2$ leads to the same answer. In this example, both eigenvalues are real and positive, so the system is not stable. It holds that $x_1(t) \to \infty$ and $x_2(t) \to -\infty$. The solution is plotted over $t \in [0, 0.4]$ in Fig. 14.1.

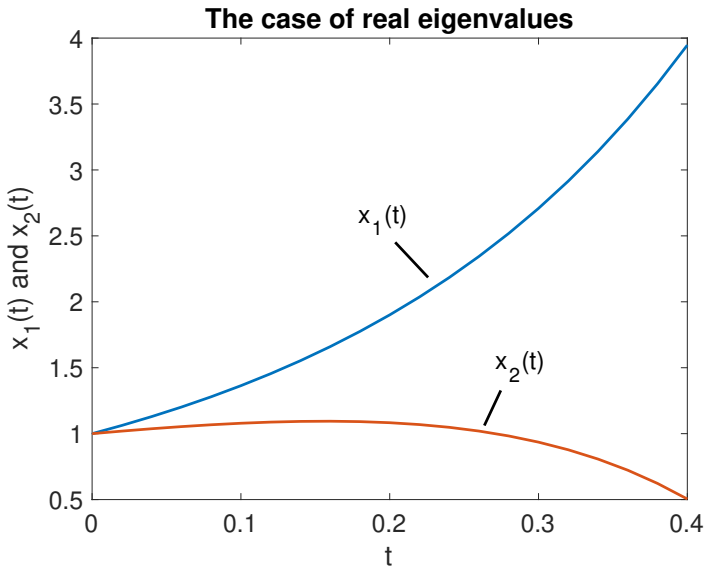


FIGURE 14.1. ODE solution $x(t) = (x_1(t), x_2(t))$ with $x(0) = (1, 1)$ when A is the matrix with real eigenvalues given in (14.6).

14.6. Example: Diagonalizable with complex eigenvalues

Diagonalizable matrices with complex eigenvalues are solved the same way. Consider

$$A = \begin{bmatrix} -3 & -2 \\ 1 & -1 \end{bmatrix} \quad , \quad x_0 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \tag{14.7}$$

To solve $x'(t) = Ax(t)$ with $x(0) = x_0$, we find the eigenvalues:

$$\det(A - \lambda I) = 0 \iff (-3-\lambda)(-1-\lambda) + 2 = 0$$

The eigenvalues are

$$\lambda_1 = -2 + i$$
$$\lambda_2 = -2 - i$$

The eigenvalues are complex conjugates of each other. These complex eigenvalues will lead us to a fully real solution. There are two distinct eigenvalues, so the matrix A is diagonalizable and has a basis of eigenvectors. To find an eigenvector for eigenvalue λ (where $\lambda = \lambda_1$ or $\lambda = \lambda_2$) we find a nontrivial solution to $(A - \lambda I)v = 0$. By Gaussian elimination we have

$$\begin{bmatrix} -3-\lambda & -2 \\ 1 & -1-\lambda \end{bmatrix} \to \begin{bmatrix} 1 & -1-\lambda \\ -3-\lambda & -2 \end{bmatrix} \to \begin{bmatrix} \boxed{1} & -1-\lambda \\ 0 & 0 \end{bmatrix}$$

where the last step uses the fact that the rows must be multiples of each other, since $A - \lambda I$ is not invertible. We have one free variable. We seek an eigenvector of the form $(a, 1)$:

$$\begin{bmatrix} 1 & -1-\lambda \\ 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ 1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \implies a + (-1-\lambda) = 0 \implies a = 1 + \lambda$$

so the eigenvectors are $v_i = (1 + \lambda_i, 1)$ for $i \in \{1, 2\}$:

$$\lambda_1 = -2 + i \implies v_1 = \begin{bmatrix} 1+\lambda_i \\ 1 \end{bmatrix} = \begin{bmatrix} -1+i \\ 1 \end{bmatrix}$$
$$\lambda_2 = -2 - i \implies v_2 = \begin{bmatrix} 1+\lambda_2 \\ 1 \end{bmatrix} = \begin{bmatrix} -1-i \\ 1 \end{bmatrix}$$

Then a general solution to $x'(t) = Ax(t)$ is

$$x(t) = \theta_1 \begin{bmatrix} -1+i \\ 1 \end{bmatrix} e^{(-2+i)t} + \theta_2 \begin{bmatrix} -1-i \\ 1 \end{bmatrix} e^{(-2-i)t} \quad \forall t \in \mathbb{R}$$

where θ_1, θ_2 are any scalar weights. In general, the solution is complex-valued, but suitable choices of θ_1, θ_2 will lead to a fully real solution. To match the desired initial condition $x(0) = (1, 3)$ we have

$$x(0) = \theta_1 \begin{bmatrix} -1+i \\ 1 \end{bmatrix} + \theta_2 \begin{bmatrix} -1-i \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

and so

$$\begin{bmatrix} -1+i & -1-i \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \implies \begin{bmatrix} \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} -1+i & -1-i \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 3 \end{bmatrix} = \begin{bmatrix} (3/2)-2i \\ (3/2)+2i \end{bmatrix}$$

Therefore

$$\boxed{x(t) = ((3/2)-2i) \begin{bmatrix} -1+i \\ 1 \end{bmatrix} e^{(-2+i)t} + ((3/2)+2i) \begin{bmatrix} -1-i \\ 1 \end{bmatrix} e^{(-2-i)t} \quad \forall t \in \mathbb{R}}$$

One can verify $x(0) = (1, 3)$ and $x'(t) = Ax(t)$. Note that $x(t)$ is a sum of two complex conjugate terms. So $x(t)$ is fully real and is equal to twice the real part of the first term. Equivalent forms of the solution are, for all $t \in \mathbb{R}$:

$$\begin{aligned} x(t) &= 2Re\left[((3/2)-2i) \begin{bmatrix} -1+i \\ 1 \end{bmatrix} e^{(-2+i)t} \right] \\ &= Re\left[(3-4i) \begin{bmatrix} -1+i \\ 1 \end{bmatrix} e^{-2t}(\cos(t) + i\sin(t)) \right] \\ &= e^{-2t} \begin{bmatrix} \cos(t) - 7\sin(t) \\ 3\cos(t) + 4\sin(t) \end{bmatrix} \end{aligned}$$

$$x(t) = \underbrace{\begin{bmatrix} -1+i & -1-i \\ 1 & 1 \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} e^{(-2+i)t} & 0 \\ 0 & e^{(-2-i)t} \end{bmatrix}}_{e^{\Lambda t}} \underbrace{\begin{bmatrix} -1+i & -1-i \\ 1 & 1 \end{bmatrix}^{-1}}_{S^{-1}} \underbrace{\begin{bmatrix} 1 \\ 3 \end{bmatrix}}_{x_0}$$

The system is stable because both eigenvalues λ_1 and λ_2 have negative real part. The solution is plotted in Fig. 14.2.

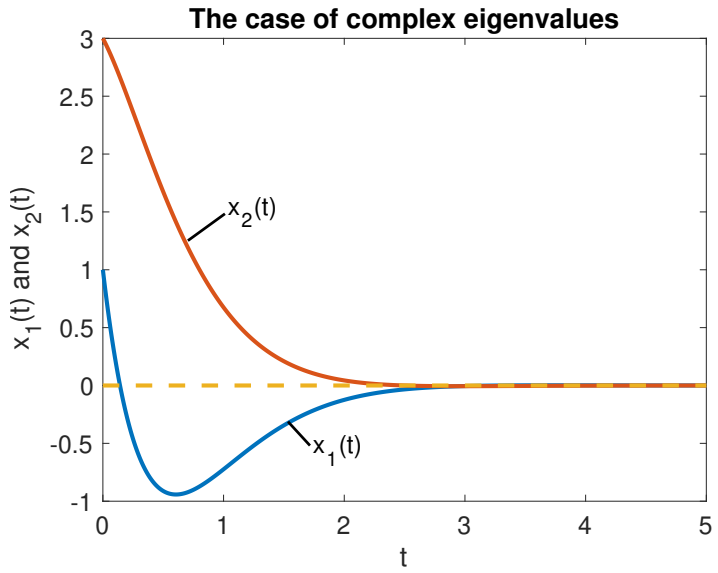


FIGURE 14.2. ODE solution $x(t) = (x_1(t), x_2(t))$ with $x(0) = (1, 3)$ when A is the matrix with complex eigenvalues given in (14.7). The real part of both eigenvalues is negative and so $x(t) \to (0, 0)$ as $t \to \infty$.

For a general real valued $n \times n$ matrix A, any complex eigenvalues must come in complex conjugate pairs (recall Chapter 13). This helps to cancel out the imaginary components to form real solutions as a sum of complex conjugate solutions. If any eigenvalues have a nonzero imaginary part, as in the example above, the corresponding real solutions to $x'(t) = Ax(t)$ typically have sine/cosine oscillations. The solution in Fig. 14.2 does not have noticeable oscillations because the oscillation frequency (determined by the imaginary part of the eigenvalues) is small in comparison to the rate of exponential decay (determined by the real part of the eigenvalues).

Let us see what happens when the imaginary part of the eigenvalues is increased from 1 to 10. Consider

$$A = \begin{bmatrix} -4 & -104 \\ 1 & 0 \end{bmatrix}, \quad x_0 = \begin{bmatrix} 1 \\ 3 \end{bmatrix} \tag{14.8}$$

By the same method used above, one can show the eigenvalues and eigenvectors are:

$$\lambda_1 = -2 + 10i \implies v_1 = \begin{bmatrix} \lambda_1 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 + 10i \\ 1 \end{bmatrix}$$

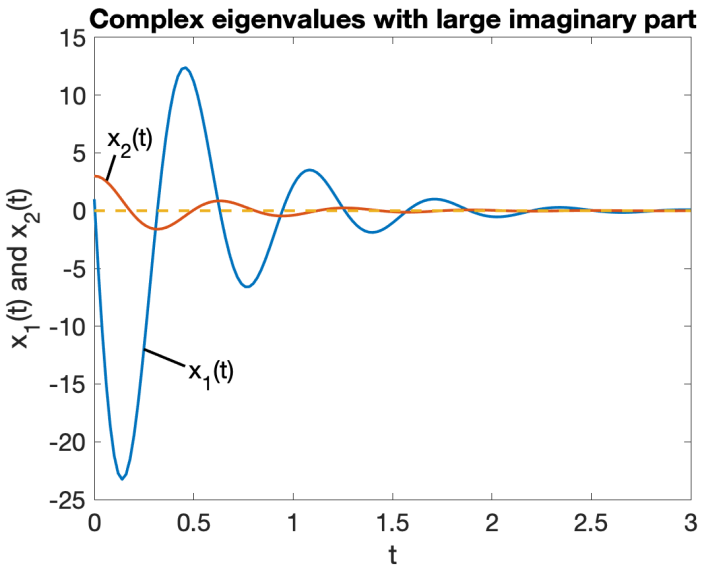


FIGURE 14.3. ODE solution $x(t) = (x_1(t), x_2(t))$ for system (14.8) where eigenvalues have an increased imaginary part in comparison to the system of Fig. 14.2. The oscillations are more apparent. The initial condition is still $x(0) = (1, 3)$. The system is still stable because the real part of the eigenvalues is still -2.

$$\lambda_2 = -2 - 10i \implies v_2 = \begin{bmatrix} \lambda_2 \\ 1 \end{bmatrix} = \begin{bmatrix} -2 - 10i \\ 1 \end{bmatrix}$$

The real part of the eigenvalues is -2, which is the same as the previous example. However, the magnitude of the imaginary part has increased from 1 to 10. The solution is

$$\begin{aligned}
x(t) &= Se^{\Lambda t}S^{-1}x_0 \\
&= \begin{bmatrix} -2+10i & -2-10i \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{(-2+10i)t} & 0 \\ 0 & e^{(-2-10i)t} \end{bmatrix} \begin{bmatrix} -2+10i & -2-10i \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 3 \end{bmatrix} \\
&= \begin{bmatrix} (1/2)+i(157/10) \\ (3/2)-i(7/20) \end{bmatrix} e^{-2t}(\cos(10t) + i\sin(10t)) + \begin{bmatrix} (1/2)-i(157/10) \\ (3/2)+i(7/20) \end{bmatrix} e^{-2t}(\cos(10t) - i\sin(10t)) \\
&= 2Re\left\{ \begin{bmatrix} (1/2)+i(157/10) \\ (3/2)-i(7/20) \end{bmatrix} e^{-2t}(\cos(10t) + i\sin(10t)) \right\} \\
&= \begin{bmatrix} 1 \\ 3 \end{bmatrix} e^{-2t}\cos(10t) + \begin{bmatrix} -157/5 \\ 7/10 \end{bmatrix} e^{-2t}\sin(10t)
\end{aligned}$$

This solution is plotted in Fig. 14.3. The system is still stable because the real part of both eigenvalues is -2. The oscillations are more significant in comparison to Fig. 14.2 because

the oscillation frequency has increased by a factor of 10. This example was created by designing a real matrix A with eigenvalues given by a desired complex conjugate pair. Exercise 14.14 shows how to design a real 2×2 matrix with eigenvalues $r + is$ and $r - is$ for any desired $r, s \in \mathbb{R}$.

14.7. Different solution structure when A is not diagonalizable

Let A be a $n \times n$ matrix (with real or complex entries). Suppose we know $\lambda \in \mathbb{C}$ is an eigenvalue of A. Generally, we want to use λ to find as many linearly independent solutions to the ODE $x'(t) = Ax(t)$ that we can. By finding a basis $\{v_1, ..., v_k\}$ for $Null(A - \lambda I)$, we have k linearly independent solutions $v_i e^{\lambda t}$. Are these the only ones we can find? This section shows how to find another solution type in the special case when

$$\dim Null[(A - \lambda I)^2] > \dim Null(A - \lambda I) \tag{14.9}$$

Let (λ, v) be an eigenpair of A. We already know that $ve^{\lambda t}$ solves the ODE. Suppose now that we can find a vector $z \in \mathbb{C}^n$ that satisfies

$$Az = \lambda z + cv \tag{14.10}$$

for some nonzero scalar $c \in \mathbb{C}$. From this information alone, it can easily be shown that (see Exercise 14.10):

- $\{v, z\}$ are linearly independent.
- $z \notin Null(A - \lambda I)$.
- $z \in Null[(A - \lambda I)^2]$.

Given $z \in \mathbb{C}^n$ that satisfies (14.10), the following function $y(t)$ solves the ODE $x'(t) = Ax(t)$ (see Exercise 14.10):

$$\boxed{y(t) = ze^{\lambda t} + ctve^{\lambda t} \quad \forall t \in \mathbb{R}}$$

So we have two *structurally different* solutions to the ODE from the same eigenvalue λ (and their initial condition vectors are linearly independent):

$$\begin{aligned} x(t) &= ve^{\lambda t} \quad , x(0) = v \\ y(t) &= ze^{\lambda t} + ctve^{\lambda t} \quad , y(0) = z \end{aligned} \tag{14.11}$$

Note that existence of the structurally different solution $y(t)$ required existence of a vector z and a scalar $c \neq 0$ that satisfy (14.10). Such z, c can be found by taking any $z \in Null[(A - \lambda I)^2]$ that is not also in $Null(A - \lambda I)$. This can be done if and only if (14.9) holds. More generally, it can be shown that A is *not diagonalizable* if and only if it has at

least one eigenvalue λ for which the condition (14.9) holds. A third type of solution can be found if $\dim Null[(A-\lambda I)^3] > \dim Null[(A-\lambda I)^2]$ (see Exercise 14.11). A full solution that works for all cases uses the *matrix exponential* developed in Section 14.8.

For example, consider

$$A = \begin{bmatrix} -1 & 2 \\ 0 & -1 \end{bmatrix} \quad , \quad x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{14.12}$$

To solve $x'(t) = Ax(t)$ with $x(0) = x_0$, we observe that the only eigenvalue is $\lambda = -1$. Also, $\dim Null(A-\lambda I) = 1$, so we have only one linearly independent eigenvector $v = (1,0)$. Specifically

$$Null(A-\lambda I) = Span\left\{\begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}$$
$$Null[(A-\lambda I)^2] = \mathbb{C}^2$$

The condition (14.9) applies. Define $\lambda = -1$, $v = (1,0)$. Then we have one solution

$$x(t) = ve^{\lambda t} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{-t} \quad \forall t$$

To fit the desired initial condition $x(0) = (1,1)$, we need another linearly independent solution. We need a vector z that satisfies (14.10). It suffices to choose any $z \in Null[(A-\lambda I)^2]$ that is not in $Null(A-\lambda I)$. That is, any $z \in \mathbb{C}^2$ that is not a multiple of $(1,0)$. For simplicity we choose $z = (0,1)$. Then

$$(A-\lambda I)z = \begin{bmatrix} 0 & 2 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 0 \end{bmatrix} = 2v$$

so equation (14.10) holds with $\lambda = -1$, $v = (1,0)$, $z = (0,1)$, $c = 2$. Using the structure (14.11), we obtain a new solution

$$y(t) = ze^{\lambda t} + ctve^{\lambda t} = \begin{bmatrix} 0 \\ 1 \end{bmatrix} e^{-t} + 2t\begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{-t} \quad \forall t$$

We obtain the desired solution $w(t)$ as a linear combination of $x(t)$ and $y(t)$:

$$w(t) = \theta_1 x(t) + \theta_2 y(t) \implies w(0) = \theta_1\begin{bmatrix} 1 \\ 0 \end{bmatrix} + \theta_2\begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

So the weights are $\theta_1 = \theta_2 = 1$. Then

$$\boxed{w(t) = x(t) + y(t) = \begin{bmatrix} 1 \\ 1 \end{bmatrix} e^{-t} + 2t \begin{bmatrix} 1 \\ 0 \end{bmatrix} e^{-t} \quad \forall t}$$

It is clear that $w(t) \to (0,0)$ as $t \to \infty$. This is because the only eigenvalue has a negative real part, so the system is stable. The solution is plotted in Fig. 14.4.

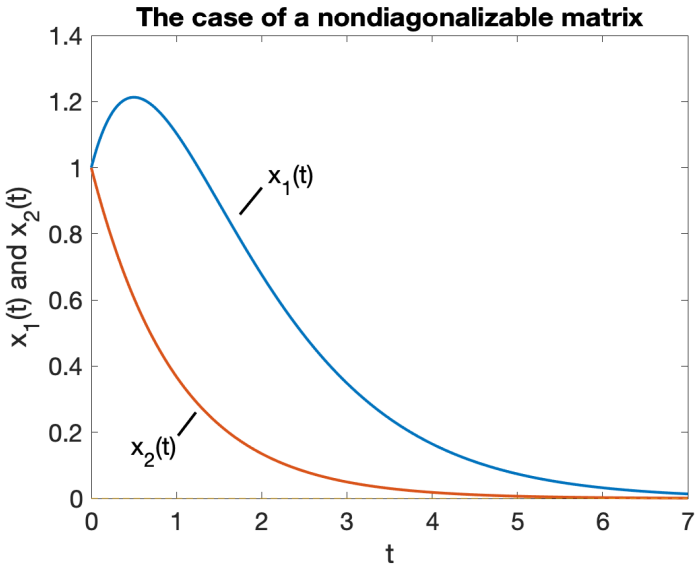


FIGURE 14.4. ODE solution $x(t) = (x_1(t), x_2(t))$ with $x(0) = (1,1)$ when A is the nondiagonalizable matrix given in (14.12). The real part of the only eigenvalue is negative and so $x(t) \to (0,0)$ as $t \to \infty$.

14.8. Matrix exponential

Let A be a $n \times n$ matrix with complex entries. We define the *matrix exponential* e^A by

$$e^A = I + \sum_{k=1}^{\infty} \frac{1}{k!} A^k$$

where I is the $n \times n$ identity matrix and

$$\sum_{k=1}^{\infty} \frac{1}{k!} A^k = \lim_{m \to \infty} \sum_{k=1}^{m} \frac{1}{k!} A^k$$

It can be shown that the right-hand-side of the above equality always converges to a finite matrix. That is, for every square matrix A with complex entries, e^A is again a square

matrix with complex entries. If all entries of A are real, then all entries of e^A are real.[1] For simplicity of notation we write

$$e^A = \sum_{k=0}^{\infty} \frac{1}{k!} A^k \tag{14.13}$$

$$e^{At} = \sum_{k=0}^{\infty} \frac{t^k}{k!} A^k \quad \forall t \in \mathbb{R} \tag{14.14}$$

These two formulas use the convention $t^0 = 1$ and $A^0 = I$ for all real numbers t and all square matrices A. Formula (14.14) views At as the matrix formed by multiplying all entries of A by the scalar t, and uses $(At)^k = t^k A^k$.

From these formulas and the fact that $I^k = I$ for all $k \in \{0, 1, 2, \ldots\}$, it is easy to see that

$$e^{It} = e^t I \quad \forall t \in \mathbb{R} \tag{14.15}$$

If 0 is the square all-zero matrix then $0^k = 0$ for $k \geq 1$, and $0^0 = I$ (by our convention in formulas (14.13) and (14.14)), so

$$(0 = \text{square all-zero matrix}) \implies e^{0t} = e^0 = I \quad \forall t \in \mathbb{R}$$

If A and B are square matrices of the same size, can we always say $e^{A+B} = e^A e^B$? No. It is typical to have

$$e^{A+B} \neq e^A e^B \neq e^B e^A$$

although these inequalities become equalities when A and B commute, as shown in the next lemma (see also Exercises 14.24 and 14.25).

LEMMA 14.1. *Let A and B be square matrices of the same size and with complex entries. Then*

$$(AB = BA) \implies e^{A+B} = e^A e^B = e^B e^A$$

PROOF. See Exercise 14.24. □

A key property of the matrix exponential is that it allows easy solutions to $x'(t) = Ax(t)$ for any square matrix A, including the case when A is not diagonalizable. Indeed, fix $x_0 \in \mathbb{C}^n$. Let A be any complex-valued $n \times n$ matrix and consider the ODE $x'(t) = Ax(t)$

[1] The infinite sum formula for e^A does not make sense when A has entries in a general field $\mathbb{F}$ because: (i) The scalars $\frac{1}{k!}$ are not necessarily defined for a general field; (ii) We do not define *convergence to a limit* for a general field. For example, in the binary field $\mathbb{F}_2 = \{0, 1\}$, finite sums make sense but infinite sums do not. While ϵ and δ proofs are common in calculus, they are not common in linear algebra. One reason is that *linear combinations of vectors always use finite sums.*

for $t \in \mathbb{R}$, with initial condition $x(0) = x_0$. The unique solution to the ODE that meets this initial condition is

$$\boxed{x(t) = e^{At}x_0 \quad \forall t \in \mathbb{R}} \tag{14.16}$$

This can be verified as follows: With $x(t)$ defined by (14.16) we have

$$x(0) = e^{A0}x_0 = Ix_0 = x_0$$

and so this function $x(t)$ satisfies the desired initial condition. Also

$$\begin{aligned} x'(t) &= \frac{d}{dt}\sum_{k=0}^{\infty}\frac{t^k}{k!}A^k x_0 \\ &= \sum_{k=1}^{\infty}\frac{kt^{k-1}}{k!}A^k x_0 \\ &= A\sum_{k=1}^{\infty}\frac{t^{k-1}}{(k-1)!}A^{k-1}x_0 \\ &= A\sum_{i=0}^{\infty}\frac{t^i}{i!}A^i x_0 \\ &= Ax(t) \end{aligned}$$

where we have differentiated term-by-term. Thus, $x'(t) = Ax(t)$ is satisfied for all $t \in \mathbb{R}$. Thus, (14.16) is indeed a solution with initial condition x_0, and the existence and uniqueness theorem ensures it is the unique solution with this initial condition.

How to compute e^{At}? Useful special cases are below (a general formula is in Section 14.9).

1. Zero matrix and identity matrix: Let 0 and I denote the $n \times n$ all-zero matrix, and identity matrix, respectively. Then

$$\begin{aligned} e^{0t} &= e^0 = I \quad \forall t \in \mathbb{R} \\ e^{It} &= e^t I \quad \forall t \in \mathbb{R} \end{aligned}$$

2. Diagonal matrix: When Λ is a diagonal matrix with diagonal entries $\lambda_1, \ldots, \lambda_n$ then

$$\Lambda^k = \begin{bmatrix} \lambda_1^k & 0 & 0 & \cdots & 0 \\ 0 & \lambda_2^k & 0 & \cdots & 0 \\ 0 & 0 & \lambda_3^k & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \lambda_n^k \end{bmatrix} \quad \forall k \in \{0, 1, 2, \ldots\}$$

and the matrix exponential formula (14.14) reduces to

$$e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & 0 & \cdots & 0 \\ 0 & 0 & e^{\lambda_3 t} & \cdots & 0 \\ \vdots & \cdots & \cdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix} \tag{14.17}$$

This is consistent with the special case $e^{It} = e^t I$ given in (14.15).

3. Diagonalizable matrix: Suppose A is a diagonalizable matrix, so $A = S\Lambda S^{-1}$ for some invertible matrix S and some diagonal matrix Λ. Then $A^k = S\Lambda^k S^{-1}$ for all $k \in \{0, 1, 2, \ldots\}$ and formula (14.14) gives

$$\boxed{\left(A = S\Lambda S^{-1}\right) \implies e^{At} = Se^{\Lambda t}S^{-1}} \tag{14.18}$$

where $e^{\Lambda t}$ has the simple form (14.17) because Λ is a diagonal matrix. This formula contains the previous examples as special cases. Of course, in this case, the ODE solution (14.16) reduces to the solution (14.3) that is already known when A is diagonalizable.

The matrix exponential e^{At} can be computed for general square matrices A, even when A is not diagonalizable, using a more complicated formula given in Section 14.9. However, another way to use the matrix exponential (for any square matrix A) is to approximate it by a finite sum:

$$e^{At} \approx \sum_{k=0}^{m} \frac{t^k}{k!} A^k$$

for some positive integer m. Then the following function $\tilde{x}(t)$ is an *approximate solution* that is accurate in some interval about $t = 0$ (with accuracy that depends on the number

of terms m in the approximation):

$$\tilde{x}(t) = \left(\sum_{k=0}^{m} \frac{t^k}{k!} A^k \right) x_0 \quad \forall t \in \mathbb{R}$$

It can be observed that $\tilde{x}(t)$ matches the desired initial condition exactly: $\tilde{x}(0) = x_0$. A disadvantage of this approach is that A^k must be computed for all $k \in \{1, \ldots, m\}$, and m should be large for accurate approximations.

Complexity note: Given $A \sim n \times n$ and given $x_0 \in \mathbb{C}^n$, an efficient way to compute $A^k x_0$ for $k \in \{1, \ldots, m\}$ is as follows:

$$\begin{aligned}
h_1 &= Ax_0 \quad [n^2 \text{ complexity to compute } h_1] \\
h_2 &= Ah_1 \quad [n^2 \text{ complexity to compute } h_2] \\
h_3 &= Ah_2 \quad [n^2 \text{ complexity to compute } h_3] \\
&\vdots \\
h_m &= Ah_{m-1} \quad [n^2 \text{ complexity to compute } h_m]
\end{aligned}$$

Thus, $h_k = A^k x_0$ for all $k \in \{1, \ldots, m\}$. Each step k multiplies the $n \times n$ matrix A with the $n \times 1$ vector h_{k-1}, so each step has complexity n^2 (measured in multiply-adds). The total complexity of m steps is then mn^2. This mn^2 complexity is the same as the complexity that would arise if, for each $k \in \{1, \ldots, m\}$, a magical genie were to give us the pre-computed $n \times n$ matrix A^k, since then we would still need to expend n^2 complexity computing $(A^k)x_0$ (and doing this m times incurs mn^2 complexity). In general, it is easier to manage the multiplication between a matrix and a vector rather than a matrix and a matrix. Indeed, multiplying two $n \times n$ matrices, such as computing A^2, has n^3 complexity. See also Exercise 2.17.

14.9. Exact solution via matrix exponential

Let A be any square matrix with complex entries. Let I be the identity matrix of the same size as A. Two useful properties of the matrix exponential are:

$$e^{cI} = e^c I \quad \forall c \in \mathbb{C} \tag{14.19}$$

$$e^{cI+A} = e^{cI} e^A \quad \forall c \in \mathbb{C} \tag{14.20}$$

Property (14.20) holds because cI and A commute, that is, $(cI)A = A(cI) = cA$ (see Exercise 14.24).

Let $\lambda_1, \ldots, \lambda_k$ be the distinct eigenvalues of A, and let m_i be the algebraic multiplicity of

eigenvalue λ_i, so

$$m_1 + \ldots + m_k = n$$

For each $i \in \{1, \ldots, k\}$ define the following subspaces of $\mathbb{C}^n$:

$$S_i = Null[(A - \lambda_i I)^n]$$

The set S_i is called the *generalized eigenspace* for eigenvalue λ_i. It can be shown that

$$m_i = \dim(S_i) \quad \forall i \in \{1, \ldots, k\}$$

As in Section 5.10, for each $i \in \{1, \ldots, k\}$ we have the following chain of nullspaces:

$$\phi \subsetneq Null(A - \lambda_i I) \subseteq Null[(A - \lambda_i I)^2] \subseteq \cdots \subseteq Null[(A - \lambda_i I)^n] \subseteq \mathbb{C}^n$$

Define d_i as the *depth* of the space S_i, defined as the first positive integer j for which

$$Null[(A - \lambda_i I)^j] = Null[(A - \lambda_i I)^{j+1}]$$

As in Section 5.10, it can be shown that for each $i \in \{1, \ldots k\}$:

$$1 \leq d_i \leq m_i \leq n$$

$$S_i = Null[(A - \lambda_i I)^j] \quad \forall j \geq d_i$$

In particular

$$(A - \lambda_i I)^j w = 0 \quad \forall w \in S_i, j \geq d_i \tag{14.21}$$

If $\{v_1^{(i)}, \ldots, v_{m_i}^{(i)}\}$ are basis vectors for S_i, then all $m_1 + \ldots + m_k = n$ vectors together form a basis for $\mathbb{C}^n$:

$$\{v_1^{(1)}, \ldots, v_{m_1}^{(1)}, \ldots, v_1^{(k)}, \ldots, v_{m_k}^{(k)}\} \text{ are a basis for } \mathbb{C}^n \tag{14.22}$$

Thus, every vector $w \in \mathbb{C}^n$ can be uniquely written as a sum of vectors in $S_1, \ldots, S_k$:

$$w = \sum_{i=1}^{k} w_i \tag{14.23}$$

where $w_i \in S_i$ for $i \in \{1, \ldots, k\}$.

To compute $e^{At}w$ for a given $w \in \mathbb{C}^n$, we first decompose w according to (14.23). Then

$$\begin{aligned}
e^{At}w &= e^{At}\sum_{i=1}^{k} w_i \\
&= \sum_{i=1}^{k} e^{At}w_i \\
&\overset{(a)}{=} \sum_{i=1}^{k} e^{\lambda_i It}e^{(A-\lambda_i I)t}w_i \\
&\overset{(b)}{=} \sum_{i=1}^{k} e^{\lambda_i t}e^{(A-\lambda_i I)t}w_i \\
&= \sum_{i=1}^{k} e^{\lambda_i t}\left(\sum_{j=0}^{\infty}\frac{t^j(A-\lambda_i I)^j w_i}{j!}\right) \\
&\overset{(c)}{=} \sum_{i=1}^{k} e^{\lambda_i t}\left(\sum_{j=0}^{d_i-1}\frac{t^j(A-\lambda_i I)^j w_i}{j!}\right)
\end{aligned}$$

where step (a) uses (14.20); (b) uses (14.19); (c) uses (14.21) and the fact $w_i \in S_i = Null[(A - \lambda_i I)^{d_i}]$. Note that $e^{At}w$ is the solution to the ODE $x'(t) = Ax(t)$ for initial condition $x(0) = w$. It can be seen that if the real part of λ_i is negative for all $i \in \{1, ..., k\}$, then $e^{At}w \to 0$ as $t \to \infty$ (regardless of the initial condition $w \in \mathbb{C}^n$). Also, this convergence to zero is exponentially fast. Thus, the ODE system $x'(t) = Ax(t)$ is said to be *stable* if all eigenvalues of A have negative real part. On the other hand, if at least one eigenvalue has *positive* real part, there exists an initial condition $w \in \mathbb{C}^n$ under which the solution $x(t)$ grows exponentially with time (possibly with sine/cosine oscillations). If at least one eigenvalue has 0 real part, there is an initial condition under which $x(t)$ stays bounded (with sine/cosine oscillations) but does not converge to zero.

Given $w \in \mathbb{C}^n$, the decomposition (14.23) is easily done using the basis vectors for each subspace S_i. Indeed, let V be the $n \times n$ matrix with columns being the basis vectors in (14.22). For notational convenience, relabel these vectors as $\{v_1, ..., v_n\}$, so

$$\begin{aligned}
S_1 &= Span\{v_1, ..., v_{m_1}\} \\
S_2 &= Span\{v_{m_1+1}, ..., v_{m_1+m_2}\} \\
&\vdots \\
S_k &= Span\{v_{m_1+...+m_{k-1}+1}, ..., v_n\}
\end{aligned}$$

Then every $w \in \mathbb{C}^n$ can be written as a linear combination of $\{v_1, \ldots, v_n\}$:

$$w = \sum_{j=1}^{n} \theta_j v_j$$

for some scalar weights $\theta = (\theta_1, \ldots, \theta_n)$. This is equivalent to $w = V\theta$, so we obtain the scalar weights $\theta = V^{-1}w$. Then the decomposition (14.23) holds for

$$\begin{aligned} w_1 &= \sum_{j=1}^{m_1} \theta_j v_j \in S_1 \\ w_2 &= \sum_{j=m_1+1}^{m_1+m_2} \theta_j v_j \in S_2 \\ &\vdots \\ w_k &= \sum_{j=m_1+\ldots+m_{k-1}+1}^{n} \theta_j v_j \in S_k \end{aligned}$$

The $n \times n$ matrix e^{At} itself can be obtained similarly (for all $t \in \mathbb{R}$):

$$e^{At} = \sum_{i=1}^{k} e^{\lambda_i t} \left(\sum_{j=0}^{d_i - 1} \frac{t^j (A - \lambda_i I)^j}{j!} \right) V_i H_i^\top \quad \forall t \in \mathbb{R}$$

where V_i is the $n \times m_i$ matrix corresponding to the block of m_i columns of the $n \times n$ matrix V pertaining to subspace S_i; $H_i^\top$ is the $m_i \times n$ matrix of the corresponding block of m_i rows of V^{-1}. This is because every $w \in \mathbb{C}^n$ is mapped to its unique ith component w_i by

$$w_i = V_i H_i^\top w$$

14.10. Discretizing for nonlinear ODEs

Nonlinear ODEs can be approximately solved by discretization: Let $f : \mathbb{C}^n \to \mathbb{C}^n$ be any (possibly nonlinear) function and consider the ODE

$$x'(t) = f(x(t)) \quad \forall t \geq 0$$

We are given an initial condition $x_0 \in \mathbb{C}^n$ and want to propagate the equation forward in time. Fix $\delta > 0$. Observe that if $\delta \approx 0$ then

$$\begin{aligned} x(t + \delta) &\approx x(t) + x'(t)\delta \\ &= x(t) + f(x(t))\delta \end{aligned}$$

This can be used as follows: Consider a fine grid of points in the timeline:

$$t \in \{0, \delta, 2\delta, 3\delta, \ldots\}$$

Then

$$\begin{aligned} x(k\delta + \delta) &\approx x(k\delta) + x'(k\delta)\delta \\ &= x(k\delta) + f(x(k\delta))\delta \end{aligned}$$

Let $y[k] \approx x(k\delta)$ be an approximation at time $k\delta$. We can precisely define the approximation sequence $\{y[k]\}_{k=0}^{\infty}$ by

$$\begin{aligned} y[k+1] &= y[k] + f(y[k])\delta \quad \forall k \in \{0, 1, 2, \ldots\} \\ y[0] &= x_0 \end{aligned}$$

The approximation is good when δ is small. When the ODE is linear, so $f(x) = Ax$ for some $n \times n$ matrix A, then we can compare the approximation to the exact solution:

- ODE:

  $$\begin{aligned} x'(t) &= Ax(t) \quad \forall t \geq 0 \\ x(0) &= x_0 \end{aligned}$$

  with exact solution $x(t) = e^{At}x_0$ for $t \geq 0$.
- Approximation with stepsize parameter $\delta > 0$ (with $y[k] \approx x(k\delta)$):

  $$\begin{aligned} y[k+1] &= (I + \delta A)y[k] \quad \forall k \in \{0, 1, 2, \ldots\} \\ y[0] &= x_0 \end{aligned}$$

  with exact solution $y[k] = (I + \delta A)^k x_0$ for all $k \in \{0, 1, 2, \ldots\}$. The sequence $\{y[k]\}_{k=0}^{\infty}$ is easily obtained by left-multiplying by $(I + \delta A)$, with no need to compute eigenvalues of A.

For example, consider

$$A = \begin{bmatrix} 1 & 4 \\ -1 & 1 \end{bmatrix} \quad , \quad x_0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \tag{14.24}$$

Then $y[k] = (I + \delta A)^k x_0$ for all $k \in \{0, 1, 2, \ldots\}$. To find $x(t) = e^{At}x_0$ for all t, observe that A

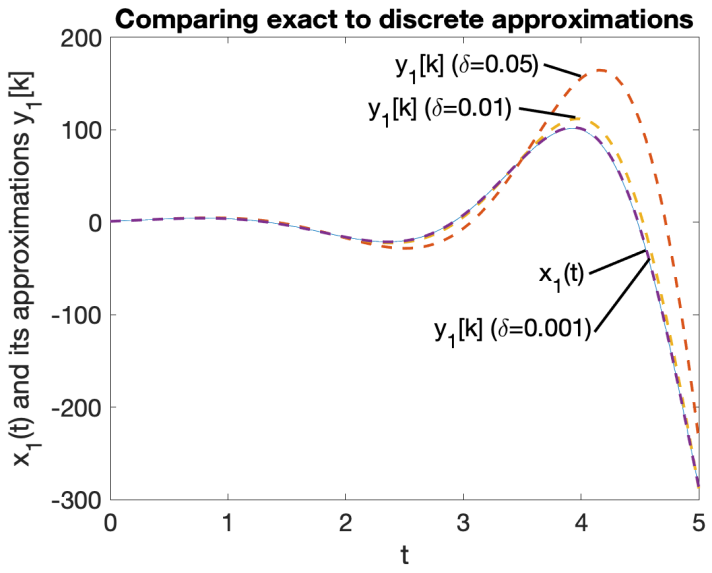


FIGURE 14.5. Comparing $x_1(t)$, which is the first component of the exact solution $x(t) = (x_1(t), x_2(t))$, to the discrete approximation $y_1[k]$ for $\delta \in \{0.05, 0.01, 0.001\}$ for the matrix A in (14.24). We use $k\delta = t$. When $\delta = 0.001$ the discrete approximation is indistinguishable from the exact solution. Note that $x_1(0) = y_1[0] = 1$ (this is not apparent in the figure because $x_1(t)$ varies between -300 and 200 over the interval plotted). The function $x_1(t)$ takes values with such large magnitudes, even though the initial condition is only $x(0) = (1, 1)$, because the system is unstable.

is diagonalizable and $A = S\Lambda S^{-1}$ where

$$S = \begin{bmatrix} -2i & 2i \\ 1 & 1 \end{bmatrix} \quad , \quad \Lambda = \begin{bmatrix} 1+2i & 0 \\ 0 & 1-2i \end{bmatrix} \Longrightarrow e^{At}x_0 = Se^{\Lambda t}S^{-1}x_0 \quad \forall t$$

That is

$$\begin{aligned} e^{At}x_0 &= \begin{bmatrix} -2i & 2i \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{(1+2i)t} & 0 \\ 0 & e^{(1-2i)t} \end{bmatrix} \begin{bmatrix} -2i & 2i \\ 1 & 1 \end{bmatrix}^{-1} \begin{bmatrix} 1 \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} -2i & 2i \\ 1 & 1 \end{bmatrix} \begin{bmatrix} e^{(1+2i)t} & 0 \\ 0 & e^{(1-2i)t} \end{bmatrix} \begin{bmatrix} 1/2 + i/4 \\ 1/2 - i/4 \end{bmatrix} \\ &= \begin{bmatrix} 1/2 - i \\ 1/2 + i/4 \end{bmatrix} e^{(1+2i)t} + \begin{bmatrix} 1/2 + i \\ 1/2 - i/4 \end{bmatrix} e^{(1-2i)t} \\ &= 2Re\left\{ \begin{bmatrix} 1/2 - i \\ 1/2 + i/4 \end{bmatrix} e^{(1+2i)t} \right\} \end{aligned}$$

$$= e^t Re\left\{ \begin{bmatrix} 1-2i \\ 1+i/2 \end{bmatrix} (\cos(2t) + i\sin(2t)) \right\}$$
$$= e^t \begin{bmatrix} \cos(2t) + 2\sin(2t) \\ \cos(2t) - (1/2)\sin(2t) \end{bmatrix}$$

The system is unstable because the eigenvalues have positive real parts. Fig. 14.5 compares the true solution $x(t) = (x_1(t), x_2(t))$ against approximations $y^{(\delta)}[k] = (y_1^{(\delta)}[k], y_2^{(\delta)}[k])$ for the three cases $\delta = 0.05$, $\delta = 0.01$, $\delta = 0.001$. Time is plotted using the conversion $t = \delta k$. For simplicity, the figure plots only the first component of $x(t)$ and $y[k]$ ($x_1(t)$ and $y_1[k]$). The approximation is accurate over a finite interval about 0 when δ is small. In the example, the approximation is indistinguishable from the exact solution when $\delta = 0.001$. While $x_2(t)$ and $y_2[k]$ are not shown, their comparison is qualitatively similar.

14.11. Linear feedback control

An important application of linear ODE theory is the design of a stabilizing controller for a physical system. Suppose we have a system with state $x(t) \in \mathbb{R}^n$ that evolves according to some ODE. Often, we can affect the dynamics of the ODE by adding a *control input signal* $u(t)$. Let $u(t) \in \mathbb{R}^m$ be the control input signal and assume the overall system is mathematically modeled with the following ODE structure:

$$x'(t) = Ax(t) + Bu(t) \quad \forall t \in \mathbb{R} \tag{14.25}$$

with given initial condition $x(0) = x_0 \in \mathbb{R}^n$. Here, A is a given $n \times n$ matrix and B is a given $n \times m$ matrix. The matrix A describes how the current system state $x(t)$ affects its own rate of change. The matrix B describes how the vector control signal $u(t)$ affects the rate of change of the $x(t)$ state. The controller is allowed to use any $u(t)$ function as an input to this system.

The mathematical model (14.25) is useful when our physical system can be described by differential equations. Typically, we want our physical system to have a specific vector state, and $x(t)$ represents the *deviation* from this desired state. Then we want $x(t)$ to be close to zero in all components and for all time. However, the system can experience disturbances, and we can view x_0 as a nonzero initial condition that occurs because of a disturbance. Alternatively, there is no disturbance, but x_0 is simply a nonzero initial deviation from the desired state. We want to design a signal $u(t)$ that drives $x(t)$ to near-zero in all components.

One way to choose a control input is to choose $u(t)$ as a linear function of the current

$x(t)$ state:

$$u(t) = -Kx(t) \quad \forall t$$

where K is a given $m \times n$ matrix that must be designed. This is a type of *linear feedback control*. It uses the current state $x(t)$ as feedback to direct the control input. Using $-K$ instead of K is a standard notational choice that reminds us that we are pushing back against the current state to drive future states toward zero. The control input structure $u(t) = -Kx(t)$ is useful because it reduces design of the system to the design of a $m \times n$ matrix K. Also, for a given matrix K, the system (14.25) reduces to the following linear ODE:

$$x'(t) = (A - BK)x(t)$$

Given an initial condition $x_0 \in \mathbb{R}^n$, the solution is $x(t) = e^{(A-BK)t}x_0$ for all $t \geq 0$. A standard goal is to design a matrix K so that all eigenvalues of $A - BK$ have negative real part, which means the ODE is stable and $x(t) \to 0$ (and convergence is exponentially fast). It may or may not be possible to design such a matrix K. The theory of such systems is extensive and is treated in more detail in control theory courses. One simple example problem of this type (when A and B are 2×2 matrices) is in Exercise 14.20.

We can use a control input of the type $u(t) = -Kx(t)$ when the controller can directly observe *all components* of the $x(t)$ vector. However, in many cases, the controller only knows a few of the components of $x(t)$. In these cases of *partial state knowledge* we define $y(t) \in \mathbb{R}^k$ as a vector of observations about the system state $x(t)$:

$$y(t) = Cx(t)$$

for some given $k \times n$ matrix C. For example, if $x(t) = (x_1(t), x_2(t), x_3(t))$ but the system controller can only observe $x_1(t)$ and $x_2(t)$, then $y(t) = (x_1(t), x_2(t))$ and

$$y(t) = \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}}_{C} \begin{bmatrix} x_1(t) \\ x_2(t) \\ x_3(t) \end{bmatrix}$$

In this example, C is a 2×3 matrix. If all three components of $x(t)$ can be observed, we could use C as the 3×3 identity matrix (which puts us back in the full information scenario $y(t) = x(t)$). In more general settings, our observations might be linear combinations of the components of $x(t)$. In practice, the observation vector $y(t)$ is a vector of measurements from sensors that monitor the system. The model $y(t) = Cx(t)$ is useful when sensor measurements are linear in the system state vector $x(t)$. Is it possible to stabilize the ODE by only observing $y(t)$, without directly knowing the state vector $x(t)$?

The mathematical formulation of such a problem is the ODE system

$$x'(t) = Ax(t) + Bu(t) \quad \forall t \in \mathbb{R} \tag{14.26}$$

$$y(t) = Cx(t) \quad \forall t \in \mathbb{R} \tag{14.27}$$

where A is a $n \times n$ matrix that describes the physical system; B is a $n \times m$ matrix that describes how our control signal $u(t)$ interacts with the physical system; C is a $k \times n$ matrix that describes our sensor measurements as linear combinations of $x(t)$. Now, we can consider linear feedback controllers of the form

$$u(t) = -Ky(t)$$

for some matrix K that we design. In particular, we base our control input $u(t)$ on a linear function of the *observation vector* $y(t)$ (which is known), rather than a linear function of the state vector $x(t)$ (which may be unknown). Here, K is a $m \times k$ matrix to be designed to stabilize the system (if possible). The system (14.26)-(14.27) then reduces to

$$x'(t) = (A - BKC)x(t) \tag{14.28}$$

Given initial condition $x_0 \in \mathbb{R}^n$, the solution is $x(t) = e^{(A-BKC)t}x_0$ for $t \geq 0$. We are given the matrices A, B, C, and we want to design a matrix K such that all eigenvalues of $(A - BKC)$ have negative real part (if it is possible to do so). A small problem of this type is in Exercise 14.21.

In the special case when we have no sensor measurements, then we have no ability to use feedback control. In this case we have $C = 0$ and the system (14.28) reduces to $x'(t) = Ax(t)$. Then stability of the system depends entirely on whether or not all eigenvalues of A have negative real part. Similarly, if the B matrix is all-zero, then (regardless of C) we have no ability to apply a control signal $u(t)$ and again the system reduces to $x'(t) = Ax(t)$. This illustrates the point that it is not always possible to stabilize the system via linear feedback control. Conditions on the matrices A, B, C that ensure existence of a matrix K that makes the ODE (14.28) stable are developed in the field of control theory. This relates to deep control theoretic concepts of *controllability* and *observability*. Exercises 14.20 and 14.21 are designed to help the reader appreciate these topics of linear feedback control. They treat small systems that can be solved using only the basic results and principles in this section.

14.12. Exercises

EXERCISE 14.1. *(real eigenvalues) Solve the ODE* $x'(t) = Ax(t)$ *for*

$$A = \begin{bmatrix} 1 & 3 \\ 5 & 1 \end{bmatrix}$$

Are solutions stable? (Examine the real part of the eigenvalues.)

EXERCISE 14.2. *(real eigenvalues) Solve the ODE* $x'(t) = Ax(t)$ *for*

$$A = \begin{bmatrix} -10 & 3 \\ 5 & -10 \end{bmatrix}$$

Are solutions stable? (Examine the real part of the eigenvalues.)

EXERCISE 14.3. *(complex eigenvalues) Solve the ODE* $x'(t) = Ax(t)$ *for*

$$A = \begin{bmatrix} 1 & 7 \\ -8 & 0 \end{bmatrix}$$

Are solutions stable? (Examine the real part of the eigenvalues.)

EXERCISE 14.4. *(complex eigenvalues) Solve the ODE* $x'(t) = Ax(t)$ *for*

$$A = \begin{bmatrix} -1 & 7 \\ -8 & -2 \end{bmatrix}$$

Are solutions stable? (Examine the real part of the eigenvalues.)

EXERCISE 14.5. *Consider the ODE* $x'(t) = Ax(t)$

$$A = \begin{bmatrix} -2 & 0 & 1 \\ 0 & -1 & 2 \\ 0 & 0 & -4 \end{bmatrix}$$

a) Before solving, will solutions be stable? (Examine the real parts of the eigenvalues.)

b) Solve the ODE.

EXERCISE 14.6. *Consider the ODE* $x'(t) = Ax(t)$

$$A = \begin{bmatrix} -2 & 5 & 0 \\ 3 & -1 & 0 \\ 0 & 0 & -1 \end{bmatrix}$$

a) Solve the ODE. Are solutions stable? (Examine the real parts of the eigenvalues.)

b) Find a nonzero initial condition for $x(0)$ *such that* $\lim_{t\to\infty} x(t) = 0$.

EXERCISE 14.7. *Consider the ODE* $x'(t) = Ax(t)$

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

a) Solve the ODE. Are solutions stable? (Examine the real parts of the eigenvalues.)

b) Find a nonzero initial condition for $x(0)$ *such that* $\lim_{t\to\infty} x(t) = 0$.

EXERCISE 14.8. *Suppose* A *is a real* $n \times n$ *matrix. Let* $x(t)$ *be a function that takes values in* $\mathbb{C}^n$ *for all time* t. *Let* $w(t), z(t)$ *be the real and imaginary parts, so*

$$x(t) = w(t) + iz(t) \quad \forall t \in \mathbb{R}$$

where $w(t), z(t) \in \mathbb{R}^n$ *for all* t.

a) Show that $x(t)$ *solves the ODE*

$$x'(t) = Ax(t) \quad \forall t \in \mathbb{R}$$

if and only if its real and imaginary parts both solve the ODE.

b) Give a counter-example to part (a) when A *is not a real matrix.*

EXERCISE 14.9. *Suppose* A *is a real* $n \times n$ *matrix. Let* $x_1(t), \ldots, x_k(t)$ *be a collection of solutions to the ODE* $x'(t) = Ax(t)$. *We want to show that the solutions* $\{x_1(t), \ldots, x_k(t)\}$ *are linearly independent (as* $\mathbb{C}^n$*-valued functions of* t*) if and only if their initial conditions* $\{x_1(0), \ldots, x_k(0)\}$ *are linearly independent (as vectors in* $\mathbb{C}^n$*).*

a) Show that if $\{x_1(0), \ldots, x_k(0)\}$ *are linearly independent, then* $\{x_1(t), \ldots, x_k(t)\}$ *are linearly independent.*

b) Show that if $\{x_1(t), \ldots, x_k(t)\}$ *are linearly independent, then* $\{x_1(0), \ldots, x_k(0)\}$ *are linearly independent. Hint: This part requires the existence and uniqueness theorem for the ODE* $x'(t) = Ax(t)$.

c) To emphasize that part (b) uses the ODE structure, give an example of two vector-valued

functions $\{x(t), y(t)\}$ that take values in $\mathbb{R}^2$ and that are linearly independent (as functions of time), but $\{x(0), y(0)\}$ are not linearly independent as two vectors in $\mathbb{R}^2$.

EXERCISE 14.10. *(A general solution when A is not diagonalizable) Suppose A is a real $n \times n$ matrix. Suppose (λ, v) is an eigenpair of A. Suppose z is a vector such that*

$$Az = \lambda z + cv$$

for some scalar $c \neq 0$.

a) Show that $\{v, z\}$ are linearly independent.

b) Show that $z \notin Null(A - \lambda I)$ but $z \in Null[(A - \lambda I)^2]$.

c) Define $y(t) = ze^{\lambda t} + rtve^{\lambda t}$ for all $t \in \mathbb{R}$, where r is some particular scalar. Find a scalar value r for which $y(t)$ solves the ODE $x'(t) = Ax(t)$.

EXERCISE 14.11. *(A solution from $Null[(A-\lambda I)^3]$)) Suppose A is a real $n \times n$ matrix. Let (λ, v) be an eigenpair of A. Suppose z_1, z_2 are vectors such that:*[2]

$$Az_1 = \lambda z_1 + c_1 v$$
$$Az_2 = \lambda z_2 + c_2 z_1$$

for some nonzero scalars c_1, c_2 (so $c_1 \neq 0$ and $c_2 \neq 0$).

a) Show that $v \in Null(A - \lambda I)$, $z_1 \in Null[(A - \lambda I)^2]$, $z_2 \in Null[(A - \lambda I)^3]$.

b) Show that $\{v, z_1, z_2\}$ are linearly independent.

c) Find scalars r_1, r_2 (in terms of c_1, c_2) such that the following function solves the ODE $x'(t) = Ax(t)$:

$$y(t) = z_2 e^{\lambda t} + r_1 t z_1 e^{\lambda t} + r_2 t^2 v e^{\lambda t} \quad \forall t$$

EXERCISE 14.12. *(Matrix exponential for diagonalizable A) Let A be a diagonalizable $n \times n$ matrix with complex entries, so $A = S\Lambda S^{-1}$ for some $n \times n$ matrices S, Λ, where S is invertible and Λ is diagonal. Show that*

$$e^{At} = Se^{\Lambda t}S^{-1} \quad \forall t \in \mathbb{R}$$

where

$$e^{\Lambda t} = \begin{bmatrix} e^{\lambda_1 t} & 0 & 0 & \cdots & 0 \\ 0 & e^{\lambda_2 t} & 0 & \cdots & 0 \\ 0 & 0 & e^{\lambda_3 t} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & e^{\lambda_n t} \end{bmatrix}$$

[2] It can be shown that such vectors z_1, z_2 exist if and only if $Null[(A - \lambda I)^2] \neq Null[(A - \lambda I)^3]$, a condition that can only hold if A is *not* diagonalizable.

EXERCISE 14.13. *Let A be a $n \times n$ matrix with complex entries. Consider the ODE $x'(t) = Ax(t)$. Let V be the complex vector space of all differentiable functions $x(t) \in \mathbb{C}^n$. Let S be the subset of V consisting of all functions that satisfy the ODE. Show that S is a subspace of V.*

EXERCISE 14.14. *(Design a matrix with eigenvalues)*

a) Design a 2×2 matrix A with real entries and eigenvalues $-5 + 12i$ and $-5 - 12i$. Use the structure

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

b) Design a 2×2 matrix B with eigenvalues $-5 + 12i$ and $-5 - 12i$ and with all entries real and nonzero. Hint: PAP^{-1} preserves the eigenvalues of A.

c) Fix $r, s \in \mathbb{R}$ with $s \neq 0$. Design a 2×2 matrix A with real entries and eigenvalues $r + si$ and $r - si$.

EXERCISE 14.15. *(3×3 real matrix with complex eigenvalues) Design a 3×3 matrix with real entries and with eigenvalues $\lambda_1 = -1 + 3i$, $\lambda_2 = -1 - 3i$, $\lambda_3 = -2$. Hint: Use the technique of Exercise 14.14 and a block diagonal structure.*

EXERCISE 14.16. *(4×4 real matrix with complex eigenvalues) Design a 4×4 matrix with real entries and with eigenvalues $\lambda_1 = 9 + i$, $\lambda_2 = 9 - i$, $\lambda_3 = -2 + 7i$, $\lambda_4 = -2 - 7i$. Hint: Use the technique of Exercise 14.14 and a block diagonal structure.*

EXERCISE 14.17. *(5×5 real matrix with complex eigenvalues) We want a 5×5 matrix with real entries and with eigenvalues $\lambda_1 = 1 + 2i$, $\lambda_2 = 1 - 2i$, $\lambda_3 = 7$, $\lambda_4 = \lambda_5 = 1$.*

a) Design such a matrix that is diagonalizable. Hint: Use the technique of Exercise 14.14 and a block diagonal structure.

a) Design such a matrix that is not diagonalizable.

EXERCISE 14.18. *(Are these eigenvalues possible?)*

a) Can we design a 2×2 matrix A with real entries and eigenvalues $1 + i$ and $3 - 2i$?

b) Can we design a 2×2 matrix A with complex entries and eigenvalues $1 + i$ and $3 - 2i$?

EXERCISE 14.19. *(Are these eigenvalues possible for a 3×3?) Is it possible to design a 3×3 matrix A with real entries and with eigenvalues $\lambda_1 = 8$, $\lambda_2 = 1 + 4i$, $\lambda_3 = 2 + 4i$?*

EXERCISE 14.20. *(Feedback control) An important control problem is to design a vector-valued input $u(t)$ for the ODE*

$$x'(t) = Ax(t) + Bu(t)$$

where A, B are given $n \times n$ matrices. Assuming all components of $x(t)$ can be observed at time t, a linear feedback controller *has the form $u(t) = -Kx(t)$ for some $n \times n$ matrix K. The goal is to design a matrix K so the resulting system $x'(t) = (A - BK)x(t)$ is stable. So all eigenvalues of $A - BK$ must have negative real part. This problem treats design of real 2×2 matrices K of the*

general form

$$K = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

so we only need to consider four real-valued coefficients a, b, c, d.

a) Design a real matrix K such that all eigenvalues of $A - BK$ have negative real part when

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

b) Design a real matrix K such that all eigenvalues of $A - BK$ have negative real part when

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 1 \end{bmatrix}, \quad B = \begin{bmatrix} 0 & 0 \\ 1 & 1 \end{bmatrix}$$

c) Show that it is impossible to design a real matrix K such that all eigenvalues of $A - BK$ have negative real part when

$$A = \begin{bmatrix} -1 & 3 \\ 0 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

EXERCISE 14.21. *(Feedback control with partial information) Given an $n \times n$ matrix A and an $n \times m$ matrix B, consider the problem of designing a vector-valued control input $u(t)$ for the system $x'(t) = Ax + Bu(t)$ when, at each time t, we do not have full access to the vector $x(t)$. Rather, we only have the* observation *$y(t) = Cx(t)$ for some $m \times n$ matrix C. A linear controller uses $u(t) = -Ky(t)$ for some $m \times m$ matrix K. The goal is to design a real matrix K so that the resulting ODE $x'(t) = (A - BKC)x(t)$ is stable, so all eigenvalues of $A - BKC$ have negative real part. Design such a K when*

$$A = \begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}, \quad B = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, C = \begin{bmatrix} 1 & 0 \end{bmatrix}$$

[Note: In this case, the observation $y(t)$ is a scalar equal to $y(t) = x_1(t)$. So we want to stabilize the system without observing $x_2(t)$. In this case the matrix K is a 1×1 matrix, so we can call it a scalar k. So the question is really: Find a $k \in \mathbb{R}$ such that all eigenvalues of $A - kBC$ have negative real part.]

EXERCISE 14.22. *(Eigenvalues and block triangular form) Let $\mathbb{F}$ be a field. Fix n, k as positive*

integers. Consider a $(n+k)\times(n+k)$ *matrix* D *of the form*

$$D = \begin{bmatrix} A & B \\ 0 & C \end{bmatrix}$$

where $A \sim n \times n$, $B \sim n \times k$, $C \sim k \times k$, *all entries are in* $\mathbb{F}$. *Suppose* A *has eigenvalue* λ *and* C *has eigenvalue* μ.

a) Show that both λ *and* μ *are eigenvalues of* D.

b) Find an eigenvector v *of* D *with eigenvalue* λ. *Hint: Consider the form* $v = (x, 0)$.

c) Suppose μ *is* not *also an eigenvalue of* A. *Find an eigenvector of* D *with eigenvalue* μ. *Hint: Consider the form* $v = (x, y)$. *Where do you use the fact that* μ *is* not *an eigenvalue of* A?

d) Use $\mathbb{F} = \mathbb{R}$ *and*

$$A = \begin{bmatrix} 4 & 3 & 0 \\ 0 & 2 & 1 \\ 0 & 0 & 1 \end{bmatrix}, B = \begin{bmatrix} 8 & -7 \\ 1 & 2 \\ 4 & 5 \end{bmatrix}, C = \begin{bmatrix} 3 & 4 \\ -1 & 8 \end{bmatrix}$$

Then A *has eigenvalues* $4, 2, 1$ *and* C *has an eigenvalues* 4 *and* 7. *Give eigenvectors of* D *with eigenvalues* $4, 2, 1, 7$.

EXERCISE 14.23. *(Random matrix) Let* A *be a* $n \times n$ *matrix with real entries. Let* Z *be a Gaussian random variable with zero mean and variance 1. Let* I *denote the* $n \times n$ *identity matrix. Then* ZI *is the* $n \times n$ *diagonal matrix with all diagonal entries being the random variable* Z. *What is the probability that* $A - ZI$ *just so happens to be invertible?*

EXERCISE 14.24. *(*$e^{A+B} = e^A e^B$ *whenever* $AB = BA$*) Suppose* A *and* B *are square* $n \times n$ *matrices. If* A *and* B *commute, so that* $AB = BA$, *it holds that*

$$(A+B)^k = \sum_{i=0}^{k} \binom{k}{i} A^{k-i} B^i \quad \forall k \in \{0, 1, 2, \ldots\} \tag{14.29}$$

The above formula uses $C^0 = I$ *for any square matrix* C. *Hence, for* $k = 0$ *the formula* (14.29) *reduces to* $(A+B)^0 = I$, *and for* $k = 1$ *it reduces to* $A + B = A + B$. *For* $k \geq 2$, *the formula* (14.29) *does not necessarily hold when* A *and* B *do not commute.*

a) Give a specific 2×2 *example such that* $(A+B)^2 \neq A^2 + 2AB + B^2$.

b) Suppose A *and* B *are* $n \times n$ *matrices that commute. Show that*

$$(A+B)^2 = A^2 + 2AB + B^2$$
$$(A+B)^3 = A^3 + 3A^2B + 3AB^2 + B^3$$

which verifies the formula (14.29) *for the cases* $k = 2$ *and* $k = 3$.

c) Suppose A and B are $n \times n$ matrices that commute, so formula (14.29) *holds. Suppose A and B have complex entries. Fill in the missing parts:*

$$\begin{aligned}
e^{A+B} &= \sum_{k=0}^{\infty} \frac{1}{k!}(A+B)^k \\
&= \sum_{k=0}^{\infty} \frac{1}{k!} \sum_{i=0}^{k} (??) \\
&= \sum_{k=0}^{\infty} \sum_{i=0}^{k} \frac{A^{k-i}}{(k-i)!} \frac{B^i}{i!} \\
&= \sum_{i=0}^{\infty} \sum_{k=i}^{\infty} \frac{A^{k-i}}{(k-i)!} \frac{B^i}{i!} \\
&\overset{(*)}{=} \sum_{i=0}^{\infty} \sum_{u=0}^{\infty} (??) \frac{B^i}{i!} \\
&= \left(\sum_{i=0}^{\infty} ?? \right) \left(\sum_{u=0}^{\infty} ?? \right) \\
&= e^A e^B
\end{aligned}$$

where in equality (*) we have used the change of variables u =??.

d) Find e^{5I+3A} for a general square matrix A with complex entries. Your answer should be in terms of the matrix e^{3A}.

EXERCISE 14.25. *($e^{A+B} \neq e^A e^B$) Define the following diagonalizable matrices A and B by*

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 2 \end{bmatrix} \quad , \quad B = \begin{bmatrix} 1 & 3 \\ 0 & 3 \end{bmatrix}$$

a) Diagonalize A and B in the forms $A = S\Lambda S^{-1}$ and $B = RDR^{-1}$.

b) Define $C = A + B$. Show that C is diagonalizable and diagonalize it to the form $C = VGV^{-1}$.

c) Compute e^A, e^B, e^C using the matrix exponential formula (14.18) *for the diagonalizable case.*

d) Show that $e^{A+B} \neq e^A e^B$, $e^{A+B} \neq e^B e^A$, and $e^A e^B \neq e^B e^A$.

CHAPTER 15.

Symmetric matrices, Hermitian matrices, Normal matrices

Consider matrices with real or complex entries. Which square matrices have an orthonormal basis of eigenvectors? This chapter answers the question by presenting the *Spectral Theorem*: A square matrix A has an orthonormal basis of eigenvectors if and only if $A^H A = AA^H$. Such matrices are called *normal matrices*. An important class of normal matrices are the real and symmetric matrices. This is useful because for *any* real $m \times n$ matrix A, the matrix $A^\top A$ is real and symmetric. In this way, the theory of normal matrices paves the way for the main result of the *next* chapter, which is that *any* (possibly nonsquare) matrix has a *singular value decomposition*.

It is important to use properties of the complex field. Thus, throughout this chapter the field is $\mathbb{F} = \mathbb{C}$, and all matrices are assumed to have entries in $\mathbb{C}$. There are two reasons for this:

- Every real number can be viewed as a complex number with zero imaginary part. Hence, a matrix with real entries is also a matrix with complex entries.
- Every $n \times n$ matrix A with complex entries has an eigenvalue $\lambda \in \mathbb{C}$ and a corresponding eigenvector $v \in \mathbb{C}^n$, so that $v \neq 0$ and $Av = \lambda v$. (In view of the first bullet, this means that if a square matrix has real entries, then it has complex entries, so it has an eigenvalue $\lambda \in \mathbb{C}$. Of course, λ might have a nonzero imaginary part.)

Throughout this chapter we use the following inner product and norm:

$$\langle x, y\rangle = x^H y \quad \forall x, y \in \mathbb{C}^n$$
$$||x||^2 = \langle x, x\rangle = x^H x \quad \forall x \in \mathbb{C}^n$$

Vectors x and y are *orthogonal* if $x^H y = 0$. If $x \in \mathbb{R}^n$ then $x^H = x^\top$ and orthogonality reduces to $x^\top y = 0$.

We use the term *real matrix* to mean a matrix with real entries. Likewise, a *real vector* is a vector in $\mathbb{R}^n$ for some positive integer n. Of course, real matrices and real vectors can be viewed as special cases of complex matrices and complex vectors. If A is a square and real matrix with a real eigenvalue λ, does that mean every corresponding eigenvector v is real? Certainly not: If we have a real eigenvector v, we can define a new eigenvector $z = iv$ (where $i = \sqrt{-1}$) and z must have at least one nonreal component. However, the next lemma shows *there exists* a real eigenvector.

LEMMA 15.1. *Suppose A is a real $n \times n$ matrix and λ is a real eigenvalue of A.*

a) Let $z \in \mathbb{C}^n$ be an eigenvector corresponding to the real eigenvalue λ. Let $x, y \in \mathbb{R}^n$ be the real and imaginary parts of z, so $z = x + iy$. Then $Ax = \lambda x$ and $Ay = \lambda y$.

b) There exists a real eigenvector $v \in \mathbb{R}^n$ with eigenvalue λ, so that $v \neq 0$ and $Av = \lambda v$.

PROOF. Since z is an eigenvector, we know $z \neq 0$. So either $x \neq 0$ or $y \neq 0$ (or both). Since $Az = \lambda z$ we have

$$A(x + iy) = \lambda(x + iy)$$

Grouping the real and imaginary parts gives

$$Ax + iAy = \lambda x + i\lambda y$$

Equating the real and imaginary parts gives two separate equations:

$$Ax = \lambda x$$
$$Ay = \lambda y$$

which proves part (a). If $x \neq 0$ then x is a real eigenvector of A corresponding to eigenvalue λ. On the other hand, if $x = 0$ then $y \neq 0$ and so y is a real eigenvector of A corresponding to eigenvalue λ. This proves part (b). □

15.1. Summary of results

Let A be a $n \times n$ matrix with complex entries.

- A is defined to be *symmetric* if $A^\top = A$.
- A is defined to be *Hermitian* if $A^H = A$ (so it is equal to its conjugate transpose).
- A is defined to be *normal* if $A^H A = AA^H$ (so it commutes with its conjugate transpose).

[The terms *Hermitian matrix* and *Hermitian operation* are distinct. The *Hermitian operation* is the complex transpose, which can be done on *any* complex matrix, while a *Hermitian matrix* is one that is unchanged by the Hermitian operation.]

Observe that

$$\{\text{Real and symmetric matrices}\} \subseteq \{\text{Hermitian matrices}\} \subseteq \{\text{Normal matrices}\}$$

so all real and symmetric matrices are Hermitian, and all Hermitian matrices are normal. We shall find that these three types of matrices are all special types of diagonalizable matrices: Not only do they have a basis of eigenvectors, but they have an *orthonormal basis* of eigenvectors. Some examples are below.

- Real and symmetric:
$$A = \begin{bmatrix} 5 & -2 & 1 \\ -2 & 3 & 4 \\ 1 & 4 & 0 \end{bmatrix} \implies A^\top = A$$
- Hermitian:
$$B = \begin{bmatrix} 8 & -2+i \\ -2-i & 7 \end{bmatrix} \implies B^H = B$$
- Normal:
$$C = \begin{bmatrix} 0 & 5 \\ -5 & 0 \end{bmatrix} \implies C^H C = CC^H$$
$$D = \begin{bmatrix} 0 & 1+i \\ -1+i & 2i \end{bmatrix} \implies D^H D = DD^H$$

This section develops and proves the following three results, which are collectively called the *Spectral Theorem*: Let A be a $n \times n$ matrix with real or complex entries. Then

1. Real and symmetric: Matrix A is real and symmetric if and only if it has *real* eigenvectors $\{v_1, \ldots, v_n\}$ that are an orthonormal basis for $\mathbb{R}^n$ with corresponding *real* eigenvalues $\lambda_1, \ldots, \lambda_n$. Such a matrix is diagonalizable and has the structure
$$A = S\Lambda S^\top = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$$

where S, Λ are $n \times n$ matrices, Λ is a diagonal matrix consisting of the real eigenvalues $\lambda_1, ..., \lambda_n$, columns of S are the real orthonormal eigenvectors $\{v_1, ..., v_n\}$, and $S^{-1} = S^\top$.

2. Hermitian: Matrix A is Hermitian if and only if it has (possibly complex) eigenvectors $\{v_1, ..., v_n\}$ that are an orthonormal basis for $\mathbb{C}^n$ with corresponding *real* eigenvalues $\lambda_1, ..., \lambda_n$. Such a matrix is diagonalizable and has the structure

$$A = S\Lambda S^H = \sum_{i=1}^{n} \lambda_i v_i v_i^H$$

where S, Λ are $n \times n$ matrices, Λ is the diagonal matrix of real eigenvalues $\lambda_1, ..., \lambda_n$, columns of S are the orthonormal eigenvectors $\{v_1, ..., v_n\}$, and $S^{-1} = S^H$.

3. Normal: Matrix A is normal if and only if it has (possibly complex) eigenvectors $\{v_1, ..., v_n\}$ that form an orthonormal basis for $\mathbb{C}^n$ with corresponding (possibly complex) eigenvalues $\lambda_1, ..., \lambda_n$. Such a matrix is diagonalizable and has the structure

$$A = S\Lambda S^H = \sum_{i=1}^{n} \lambda_i v_i v_i^H$$

where S, Λ are $n \times n$ matrices, Λ is the diagonal matrix of eigenvalues, columns of S are the corresponding orthonormal eigenvectors $\{v_1, ..., v_n\}$, and $S^{-1} = S^H$.

In particular, it can be shown that all Hermitian matrices have real eigenvalues (see Lemma 15.3). For key insight into the Spectral Theorem, the reader is encouraged to work on Exercises 15.5 and 15.7, which are designed to prove the following result:

LEMMA 15.2. *Suppose A is a Hermitian matrix. Let (λ_1, v_1) and (λ_2, v_2) be two eigenpairs of A. Then*

$$(\lambda_1 \neq \lambda_2) \implies v_1^H v_2 = 0$$

In other words, if $A^H = A$, then any two distinct eigenvalues of A must have orthogonal eigenvectors. In particular, this property holds for any real and symmetric matrix.

PROOF. See Exercises 15.5 and 15.7. □

15.2. Square matrices have orthonormal columns if and only if orthonormal rows

A *unitary matrix* is a square matrix with complex entries that has orthonormal columns. Let S be a $n \times n$ unitary matrix and let the orthonormal columns be $\{v_1, ..., v_n\}$, so

$$S = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

So $v_i \in \mathbb{C}^n$ for $i \in \{1, ..., n\}$ and

$$\begin{aligned} ||v_i|| = v_i^H v_i = 1 \quad \forall i \in \{1, ..., n\} \\ v_i^H v_j = 0 \quad \forall i \neq j \end{aligned}$$

Then

$$S^H S = \begin{bmatrix} v_1^H \\ v_2^H \\ \vdots \\ v_n^H \end{bmatrix} \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} = \begin{bmatrix} v_1^H v_1 & v_1^H v_2 & \cdots & v_1^H v_n \\ v_2^H v_1 & v_2^H v_2 & \cdots & v_2^H v_n \\ \vdots & \vdots & \ddots & \vdots \\ v_n^H v_1 & v_n^H v_2 & \cdots & v_n^H v_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} = I$$

Then S^H and S are square matrices that multiply to the identity matrix I, they must be inverses of each other!

$$S^{-1} = S^H$$

Since $S^{-1}S = SS^{-1}$, this implies

$$S^H S = SS^H = I$$

- The equation $S^H S = I$ means that the columns of S are orthonormal.
- The equation $SS^H = I$ means that the rows of S are orthonormal.

Therefore, if a square matrix has orthonormal columns, it must also have orthonormal rows! A similar argument shows that if a square matrix has orthonormal rows, it must also have orthonormal columns! Note that in the special case when S is a real matrix then $S^H = S^\top$. Norm-preserving properties of matrices with orthonormal columns, including the (square) unitary matrices, are explored in Exercise 8.22.

For example, consider the following real matrix with a missing column:

$$S = \begin{bmatrix} -\sqrt{2/3} & 1/4 & ? \\ \sqrt{1/3} & \sqrt{1/8} & ? \\ 0 & \sqrt{13/16} & ? \end{bmatrix} \tag{15.1}$$

The first two columns of S are orthonormal. Any way of creating a third column that is normal and orthogonal to the previous two columns must also make the three rows orthonormal. Call the first two columns v_1 and v_2, so $\{v_1, v_2\}$ are orthonormal. One way to create the third column v_3 is to let z be any vector in $\mathbb{C}^3$ that is not in $Span\{v_1, v_2\}$, strip away the projection of z onto $Span\{v_1, v_2\}$, then normalize the result:

$$q = z - (z^H v_1)v_1 - (z^H v_2)v_2$$

which means q is a nonzero vector that satisfies $q^\top v_1 = 0$, $q^\top v_2 = 0$, and so we have a third column:

$$v_3 = \frac{q}{\|q\|}$$

For example, with $z = (1, 1, 1) \notin Span\{v_1, v_2\}$ we can perform the computations in Matlab:

$$q = \begin{bmatrix} 0.4285 \\ 0.6060 \\ -0.3565 \end{bmatrix} \Longrightarrow v_3 = \begin{bmatrix} 0.5204 \\ 0.7360 \\ -0.4330 \end{bmatrix}$$

and (to within the finite decimal precision displayed) we observe that the following matrix has orthonormal rows and columns:

$$S = \begin{bmatrix} -\sqrt{2/3} & 1/4 & 0.5204 \\ \sqrt{1/3} & \sqrt{1/8} & 0.7360 \\ 0 & \sqrt{13/16} & -0.4330 \end{bmatrix}$$

Of course, another way to produce a vector in $\mathbb{C}^3$ that is orthogonal to the first two columns of S in (15.1) is to find any nonzero vector $(x, y, z) \in \mathbb{C}^3$ that solves

$$\begin{bmatrix} -\sqrt{2/3} & \sqrt{1/3} & 0 \\ 1/4 & \sqrt{2/16} & \sqrt{13/16} \end{bmatrix} \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$$

and then normalize that (x, y, z) vector, so $v_3 = (x, y, z)/\sqrt{x^2 + y^2 + z^2}$.

What if a *nonsquare* matrix A has orthonormal columns? Must it also have orthonormal rows? No. The fact that A is nonsquare with orthonormal columns means it must be a tall matrix, so we can add additional orthonormal columns to it to create a *square* matrix with orthonormal columns, and *that* square matrix must have orthonormal rows. For example consider A as the 3×2 matrix given by the first two columns of S in (15.1). That matrix A has two orthonormal columns, but its rows are neither normal nor orthogonal. Adding a third orthonormal column as we did ensures S has both orthonormal columns and orthonormal rows. On the other hand, it is still true that

$$A^\top A = \begin{bmatrix} -\sqrt{2/3} & \sqrt{1/3} & 0 \\ 1/4 & \sqrt{1/8} & \sqrt{13/16} \end{bmatrix} \begin{bmatrix} -\sqrt{2/3} & 1/4 \\ \sqrt{1/3} & \sqrt{1/8} \\ 0 & \sqrt{13/16} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Indeed, by definition of orthonormality, any real matrix A with orthonormal columns

must be either square or tall and must satisfy $A^\top A = I$. Of course, when $A^\top$ and A are not square, the fact $A^\top A = I$ does *not* imply that $A^\top$ and A are invertible. Nonsquare matrices cannot be invertible. Also, if a nonsquare $m \times n$ matrix A satisfies $A^\top A = I$ (the $n \times n$ identity), we must have $m > n$ and it is impossible for $AA^\top$ to be the $m \times m$ identity because $\text{rank}(AA^\top) \leq n$. See a related definition of *pseudoinverse* in Appendix B.

15.3. Diagonalizable form with orthonormal eigenvectors

Suppose A is a $n \times n$ matrix with real eigenvectors $\{v_1, \ldots, v_n\}$ that form an orthonormal basis for $\mathbb{R}^n$, with corresponding real eigenvalues $\lambda_1, \ldots, \lambda_n$. Thus,

$$\|v_i\|^2 = v_i^\top v_i = 1 \quad \forall i \in \{1, \ldots, n\} \tag{15.2}$$

$$v_i^\top v_j = 0 \quad \forall i \neq j \tag{15.3}$$

$$Av_i = \lambda_i v_i \quad \forall i \in \{1, \ldots, n\} \tag{15.4}$$

Define the $n \times n$ matrices Λ and S by

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \quad , \quad S = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix} \tag{15.5}$$

so Λ is a diagonal matrix of real eigenvalues; columns of S are the real eigenvectors. Orthonormality of the columns of the square matrix S implies S is unitary so $S^{-1} = S^\top$.

We claim that A must be real and symmetric and

$$A = S\Lambda S^\top = \sum_{i=1}^{n} \lambda_i v_i v_i^\top \tag{15.6}$$

The form $A = S\Lambda S^\top$ (where $S^\top = S^{-1}$). Thus, not only is A diagonalizable, but it has eigenvectors that form an orthonormal basis for $\mathbb{R}^n$. It is clear that $S\Lambda S^\top$ is real and symmetric because Λ and S are both real and

$$(S\Lambda S^\top)^\top = S\Lambda S^\top$$

Hence, if we can prove that $A = S\Lambda S^\top$, this also proves that A is real and symmetric.

The fact (15.6) is true because the three matrices

$$A \quad , \quad S\Lambda S^\top \quad , \quad \sum_{i=1}^{n} \lambda_i v_i v_i^\top$$

are $n \times n$ matrices that have the same input-output behavior on the basis $\{v_1, \ldots, v_n\}$ for $\mathbb{R}^n$, so they must all be the same matrix. To see this, fix $j \in \{1, \ldots, n\}$. We already know $Av_j = \lambda_j v_j$ by (15.4). Also

$$\left(\sum_{i=1}^{n} \lambda_i v_i v_i^\top\right) v_j = \sum_{i=1}^{n} \lambda_i v_i (v_i^\top v_j) \overset{(a)}{=} \lambda_j v_j (v_j^\top v_j) \overset{(b)}{=} \lambda_j v_j$$

where equality (a) holds by (15.3); equality (b) holds by (15.2). So we must have $A = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$.

Finally, since $S^\top S = I$ and v_j is the jth column of S, it follows that

$$S^\top v_j = e_j \tag{15.7}$$

where e_j is the jth column of I. Multiplying a $n \times n$ matrix by e_j gives the jth column and so

$$\Lambda e_j = j\text{th column of the diagonal matrix } \Lambda = \lambda_j e_j \tag{15.8}$$

Putting these equalities together gives

$$S\Lambda S^\top v_j \overset{(a)}{=} S\Lambda e_j \overset{(b)}{=} \lambda_j S e_j \overset{(c)}{=} \lambda_j v_j$$

where (a) holds by (15.7); (b) holds by (15.8); (c) holds because the jth column of S is v_j. So we must have $A = S\Lambda S^\top$.

By a similar argument, we can build a matrix A with any desired real eigenvalues $\lambda_1, \ldots, \lambda_n$ and any desired real orthonormal eigenvectors $\{v_1, \ldots, v_n\}$ by using the form $A = S\Lambda S^\top$, where Λ and S have the structure defined in (15.5). For example, consider $n = 2$ and define

$$\lambda_1 = -24, \quad \lambda_2 = 16, \quad v_1 = \begin{bmatrix} \sqrt{1/8} \\ \sqrt{7/8} \end{bmatrix}, \quad v_2 = \begin{bmatrix} -\sqrt{7/8} \\ \sqrt{1/8} \end{bmatrix}$$

Clearly $\{v_1, v_2\}$ are orthonormal. Now define

$$\Lambda = \begin{bmatrix} -24 & 0 \\ 0 & 16 \end{bmatrix}, \quad S = \begin{bmatrix} \sqrt{1/8} & -\sqrt{7/8} \\ \sqrt{7/8} & \sqrt{1/8} \end{bmatrix}, \quad S^\top = \begin{bmatrix} \sqrt{1/8} & \sqrt{7/8} \\ -\sqrt{7/8} & \sqrt{1/8} \end{bmatrix}$$

It can be verified that $S^{-1} = S^\top$. Define $A = S\Lambda S^\top$. Then

$$A = S\Lambda S^\top = \begin{bmatrix} 11 & -5\sqrt{7} \\ -5\sqrt{7} & -19 \end{bmatrix}$$

and A is indeed real, symmetric, and has eigenvectors $\{v_1, v_2\}$ with corresponding eigen-

values λ_1, λ_2. Also

$$A = \sum_{i=1}^{2} \lambda_i v_i v_i^\top = (-24) \begin{bmatrix} \sqrt{1/8} \\ \sqrt{7/8} \end{bmatrix} \begin{bmatrix} \sqrt{1/8} & \sqrt{7/8} \end{bmatrix} + (16) \begin{bmatrix} -\sqrt{7/8} \\ \sqrt{1/8} \end{bmatrix} \begin{bmatrix} -\sqrt{7/8} & \sqrt{1/8} \end{bmatrix}$$

The results of this section *start* by assuming a matrix A has an orthonormal basis of real eigenvectors with real eigenvalues, and we conclude that A must be real and symmetric. What about the other direction? If we assume a matrix is real and symmetric, does that mean it has an orthonormal basis of eigenvectors with real eigenvalues? The answer is *yes* and that is called the *Spectral Theorem*, proven in the next section.

Before proving the Spectral Theorem, we note that results of this section easily generalize to complex-valued matrices. The generalization of a symmetric matrix in the complex world is a Hermitian matrix (a matrix that satisfies $A^H = A$). An argument similar to the above can show that if A is a square matrix with orthonormal (possibly complex) eigenvectors $\{v_1, \ldots, v_n\}$ with real eigenvalues $\lambda_1, \ldots, \lambda_n$, then

$$A = S \Lambda S^H = \sum_{i=1}^{n} \lambda_i v_i v_i^H$$

where columns of S are the vectors $\{v_1, \ldots, v_n\}$ and Λ is the diagonal matrix of real eigenvalues, and so

$$A^H = (S \Lambda S^H)^H = S \Lambda^H S^H \overset{(a)}{=} S \Lambda S^H = A \tag{15.9}$$

where equality (a) uses $\Lambda^H = \Lambda$ (because Λ is a real diagonal matrix). So $A^H = A$ which means A is Hermitian. If one allows the eigenvalues to be *complex numbers* then $\Lambda^H = \Lambda$ is no longer guaranteed so the chain of equalities (15.9) stops just before (a) to conclude $A^H = S \Lambda^H S^H$, which is not enough to conclude A is Hermitian but is enough to conclude A is normal ($A^H A = A A^H$).

15.4. Spectral Theorem

Recall that a matrix A is Hermitian if $A^H = A$.

LEMMA 15.3. *If A is a Hermitian matrix, then all eigenvalues of A are real. In particular, if A is real and symmetric, then all its eigenvalues are real.*

PROOF. Suppose A is a $n \times n$ Hermitian matrix. Let $\lambda \in \mathbb{C}$ be an eigenvalue of A. We want to show λ is real. By definition of eigenvalue, there is a nonzero vector $x \in \mathbb{C}^n$ such that

$$Ax = \lambda x$$

Taking Hermitians of both sides gives

$$x^H A^H = \overline{\lambda} x^H$$

where $\overline{\lambda}$ is the complex conjugate of λ. Since $A^H = A$ the above equation reduces to

$$x^H A = \overline{\lambda} x^H$$

Multiplying both sides of the above equation on the right by x gives

$$x^H A x = \overline{\lambda} x^H x$$

Since $Ax = \lambda x$ we obtain

$$\lambda x^H x = \overline{\lambda} x^H x$$

Using $x^H x = ||x||^2$ yields

$$\lambda ||x||^2 = \overline{\lambda} ||x||^2$$

We know $||x||^2 \neq 0$ since $x \neq 0$ (it is an eigenvector). Dividing both sides by $||x||^2$ yields $\lambda = \overline{\lambda}$, so λ must be real. □

Recall that a matrix A is *normal* if $A^H A = A A^H$.

LEMMA 15.4. *If A is normal and has eigenpair (λ, v), then A^H is normal and has eigenpair $(\overline{\lambda}, v)$, where $\overline{\lambda}$ denotes the complex conjugate of λ.*

PROOF. If A is normal then $A^H A = A A^H$ and so A^H is normal. Suppose A has eigenpair (λ, v). We want to show $(\overline{\lambda}, v)$ is an eigenpair of A^H. We know

$$Av = \lambda v \tag{15.10}$$

and taking complex transposes of both sides of (15.10) gives

$$v^H A^H = \overline{\lambda} v^H \tag{15.11}$$

Taking the norm-square of both sides of (15.10) gives

$$||Av||^2 = |\lambda|^2 \cdot ||v||^2 \tag{15.12}$$

Since we know $v \neq 0$, it suffices to show $(A^H - \overline{\lambda} I)v = 0$. We have

$$\begin{aligned} \left|\left|(A^H - \overline{\lambda} I)v\right|\right|^2 &= [(A^H - \overline{\lambda} I)v]^H [(A^H - \overline{\lambda} I)v] \\ &= v^H (A - \lambda I)(A^H - \overline{\lambda} I)v \end{aligned}$$

$$
\begin{aligned}
&= v^H(AA^H - \overline{\lambda}A - \lambda A^H + |\lambda|^2 I)v \\
&\overset{(a)}{=} v^H(A^H A - \overline{\lambda}A - \lambda A^H + |\lambda|^2 I)v \\
&= (v^H A^H)(Av) - \overline{\lambda}v^H(Av) - \lambda(v^H A^H)v + |\lambda|^2 \cdot \|v\|^2 \\
&\overset{(b)}{=} (v^H A^H)(Av) - \overline{\lambda}v^H(\lambda v) - \lambda(\overline{\lambda}v^H)v + |\lambda|^2 \cdot \|v\|^2 \\
&= \|Av\|^2 - |\lambda|^2 \cdot \|v\|^2 - |\lambda|^2 \cdot \|v\|^2 + |\lambda|^2 \cdot \|v\|^2 \\
&= \|Av\|^2 - |\lambda|^2 \cdot \|v\|^2 \\
&\overset{(c)}{=} 0
\end{aligned}
$$

where (a) uses $AA^H = A^H A$; (b) substitutes (15.10) and (15.11); (c) uses (15.12). □

The following result uses the fact that if A is a $n \times n$ matrix with complex entries, and if $S \subseteq \mathbb{C}^n$ is a nontrivial subspace that is invariant under the linear map A, then A has an eigenvector in S (Theorem 13.2).

LEMMA 15.5. *Suppose A is a normal matrix with eigenpairs (λ_1, v_1), $(\lambda_2, v_2), \ldots, (\lambda_m, v_m)$ for some positive integer m. Define*

$$U = Span\{v_1, v_2, \ldots, v_m\}$$

Then the subspace $U^\perp$ is invariant under the linear map A, meaning that $w \in U^\perp \implies Aw \in U^\perp$. Further, if $\dim(U^\perp) \geq 1$ then A has a norm-1 eigenvector in $U^\perp$.

PROOF. Fix $w \in U^\perp$. It suffices to show Aw is orthogonal to v_i for all $i \in \{1, \ldots, m\}$. Fix $i \in \{1, \ldots, m\}$. We have

$$
\begin{aligned}
\langle Aw, v_i \rangle &= (Aw)^H v_i \\
&= w^H A^H v_i \\
&\overset{(a)}{=} w^H(\overline{\lambda}_i v_i) \\
&= \overline{\lambda}_i w^H v_i \\
&\overset{(b)}{=} 0
\end{aligned}
$$

where (a) holds because (λ_i, v_i) is an eigenpair of A, so Lemma 15.4 ensures $A^H v_i = \overline{\lambda}_i v_i$; (b) holds because $w \in U^\perp$ so w and v_i are orthogonal. Thus, $U^\perp$ is invariant under the linear map A. If its dimension is at least 1, it is a nontrivial invariant subspace and so it contains an eigenvector of A (Theorem 13.2). Normalizing this eigenvector completes the proof. □

THEOREM 15.1. *(Spectral Theorem) Let A be a $n \times n$ matrix with complex entries. Then:*

a) A is a normal matrix if and only if it has eigenvectors $\{v_1, ..., v_n\}$ that form an orthonormal basis for $\mathbb{C}^n$. The eigenvectors and the corresponding eigenvalues $\lambda_1, ..., \lambda_n$ are possibly complex.

b) A is a Hermitian matrix if and only if it has eigenvectors $\{v_1, ..., v_n\}$ that form an orthonormal basis for $\mathbb{C}^n$, with corresponding real *eigenvalues $\lambda_1, ..., \lambda_n$.*

c) A is a real and symmetric matrix if and only if it has real *eigenvectors $\{v_1, ..., v_n\}$ that form an orthonormal basis for $\mathbb{R}^n$, with corresponding* real *eigenvalues $\lambda_1, ..., \lambda_n$.*

PROOF. (Part (a)) Let A be a normal matrix. We know that every square matrix with complex entries has a (possibly complex) eigenvalue. Let λ_1 be this eigenvalue and let v_1 be a corresponding norm-1 eigenvector. So $||v_1|| = 1$, and $Av_1 = \lambda_1 v_1$. If $n = 1$ then $\{v_1\}$ is an orthonormal basis for $\mathbb{C}^1$ and we are done.

Now suppose $n > 1$. Let k be a positive integer in $\{1, ..., n-1\}$ and assume there are orthonormal eigenvectors $\{v_1, ..., v_k\}$ (this holds for $k = 1$). We now construct an eigenvector v_{k+1} that is orthogonal to all vectors $\{v_1, ..., v_k\}$. For this, define $U = Span\{v_1, ..., v_k\}$. Since $k < n$, the subspace $U^\perp$ has dimension at least 1 (so it is nonempty). Lemma 15.5 ensures the subspace $U^\perp$ is invariant under the linear map A and contains a norm-1 eigenvector, call this eigenvector v_{k+1}. Since $v_{k+1} \in U^\perp$, it is orthogonal to all vectors $\{v_1, ..., v_k\}$. Then $\{v_1, ..., v_{k+1}\}$ are orthonormal eigenvectors. The induction proceeds for $k \in \{1, ..., n-1\}$ until we find orthonormal eigenvectors $\{v_1, ..., v_n\}$. Since these are n linearly independent vectors in $\mathbb{C}^n$, they must form a basis for $\mathbb{C}^n$. This proves the forward direction of part (a).

Conversely, suppose A has an orthonormal basis of eigenvectors. Let S be the matrix formed by using these eigenvectors as columns, and let Λ be the diagonal matrix of eigenvalues. Then S is unitary so $S^{-1} = S^H$ and we have the diagonal form $A = S\Lambda S^H$. Taking Hermitians of both sides gives $A^H = S\Lambda^H S^H$ and it is easy to see that $A^H A = AA^H$. □

PROOF. (Part (b)) To prove this, first suppose A is Hermitian. Then it is normal and part (a) ensures it has an orthonormal basis of eigenvectors. The eigenvalues must be real because all eigenvalues of Hermitian matrices are real (Lemma 15.3). Conversely, suppose A has an orthonormal basis of eigenvectors with *real* eigenvalues. Forming S and Λ as before gives $A = S\Lambda S^H$. Taking Hermitians of both sides gives $A^H = S\overline{\Lambda}S^H$. Since Λ is real, we know $\overline{\Lambda} = \Lambda$ and so $A^H = A$. □

PROOF. (Part (c)) Suppose A is real and symmetric. We know A has at least one eigenvalue $\lambda_1 \in \mathbb{C}$. Since A is real and symmetric, we know λ_1 (and all other eigenvalues) are real. Then, Lemma 15.1 implies we can start the same induction of part (a) by using a *real* eigenvector v_1. The induction proceeds by assuming $\{v_1, ..., v_k\}$ are real orthonormal

eigenvectors and, as in part (a), finding a new eigenvector v that is orthogonal to all vectors $\{v_1, ..., v_k\}$. Both the real and imaginary parts of v must be orthogonal to all of the (real) vectors in $\{v_1, ..., v_k\}$. So we define $\tilde{v}_{k+1} \in \mathbb{R}^n$ as either the real or imaginary part of v (whichever is nonzero) and again invoke Lemma 15.1 to ensure $\tilde{v}_{k+1}$ is an eigenvector. Define the normalized vector $v_{k+1} = \tilde{v}_{k+1}/\|\tilde{v}_{k+1}\|$. Running the induction produces real and orthonormal eigenvectors $\{v_1, ..., v_n\}$, these are n linearly independent vectors in $\mathbb{R}^n$ so they must form a basis for $\mathbb{R}^n$.

Conversely, suppose A has an orthonormal basis of real eigenvectors with real eigenvalues. Define S and Λ as before, so $A = S\Lambda S^H$. However, in this case S is real and so $S^H = S^\top$. Then $A = S\Lambda S^\top$ and $A^\top = S\Lambda S^\top = A$. We know A is real because it is the product of the real matrices S, Λ, and $S^\top$. □

It is useful to remember that in all cases of the above theorem, since A has orthonormal eigenvectors $\{v_1, ..., v_n\}$ with eigenvalues $\lambda_1, ..., \lambda_n$, we can form a square matrix S with columns given by the eigenvectors, and a diagonal matrix Λ with entries given by the eigenvalues. Then S is unitary and so $S^{-1} = S^H$ and we have

$$A = S\Lambda S^H = \sum_{i=1}^{n} \lambda_i v_i v_i^H$$

In the case when A is real and symmetric, we know S is real so $S^H = S^\top$ and the formula becomes

$$A = S\Lambda S^\top = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$$

15.5. Properties of real and symmetric matrices

Let A be a real and symmetric $n \times n$ matrix. The Spectral Theorem ensures there are *real* eigenvalues $\lambda_1, ..., \lambda_n$ and orthonormal eigenvectors $\{v_1, ..., v_n\}$ that form a basis for $\mathbb{R}^n$. Define

$$\lambda_{min}(A) = \min\{\lambda_1, \ldots, \lambda_n\} \tag{15.13}$$

$$\lambda_{max}(A) = \max\{\lambda_1, \ldots, \lambda_n\} \tag{15.14}$$

$$\theta_{min}(A) = \min\{|\lambda_1|, \ldots, |\lambda_n|\} \tag{15.15}$$

$$\theta_{max}(A) = \max\{|\lambda_1|, \ldots, |\lambda_n|\} \tag{15.16}$$

That is, $\lambda_{min}(A)$ and $\lambda_{max}(A)$ are the minimum and maximum eigenvalues of the real and symmetric matrix A. Also, $\theta_{min}(A)$ and $\theta_{max}(A)$ denote the *magnitudes* of the minimum-magnitude eigenvalue and maximum-magnitude eigenvalue.

Every $x \in \mathbb{R}^n$ is a linear combination of $\{v_1, ..., v_n\}$, so

$$x = \sum_{i=1}^{n} c_i v_i \tag{15.17}$$

for some real numbers $c_1, ..., c_n$. Orthonormality of $\{v_1, ..., v_n\}$ ensures (recall Lemma 8.6b):

$$||x||^2 = \sum_{i=1}^{n} c_i^2 \tag{15.18}$$

Multiplying both sides of (15.17) by A gives

$$Ax = \sum_{i=1}^{n} c_i (Av_i) = \sum_{i=1}^{n} c_i \lambda_i v_i \tag{15.19}$$

Again, orthonormality of $\{v_1, ..., v_n\}$ implies

$$||Ax||^2 = \left|\left| \sum_{i=1}^{n} c_i \lambda_i v_i \right|\right|^2 = \sum_{i=1}^{n} c_i^2 \lambda_i^2 \tag{15.20}$$

Comparing (15.20) and (15.18) gives

$$\theta_{min}^2(A)\,||x||^2 \le ||Ax||^2 \le \theta_{max}^2(A)\,||x||^2 \quad \forall x \in \mathbb{R}^n \tag{15.21}$$

Taking square roots of (15.21) gives

$$\boxed{\theta_{min}(A)\,||x|| \le ||Ax|| \le \theta_{max}(A)\,||x|| \quad \forall x \in \mathbb{R}^n} \tag{15.22}$$

The *lower bound* in (15.22) is met with equality by any eigenvector $v_{min} \in \mathbb{R}^n$ corresponding to the minimum magnitude eigenvalue. The *upper bound* in (15.22) is met with equality by any eigenvector $v_{max} \in \mathbb{R}^n$ corresponding to the maximum magnitude eigenvalue.

We can also obtain a simple expression for $x^\top Ax$ using (15.17):

$$x^\top Ax \overset{(a)}{=} \left(\sum_{i=1}^{n} c_i v_i^\top \right) \left(\sum_{j=1}^{n} c_j \lambda_j v_j \right) = \sum_{i=1}^{n} \sum_{j=1}^{n} c_i c_j \lambda_j (v_i^\top v_j) \overset{(b)}{=} \sum_{i=1}^{n} c_i^2 \lambda_i \tag{15.23}$$

where (a) substitutes (15.17) and (15.19); (b) uses $v_i^\top v_i = 1$ and $v_i^\top v_j = 0$ for $i \neq j$ (orthonormality). Intuitively, equality (b) has filtered out terms with $i \neq j$ because $v_i^\top v_j = \delta_{ij}$, where δ_{ij} is the *Kronecker delta function* that is 0 if $i \neq j$ and 1 if $i = j$. Comparing (15.18) and (15.23)

gives

$$\boxed{\lambda_{min}(A)\,||x||^2 \le x^\top A x \le \lambda_{max}(A)\,||x||^2 \quad \forall x \in \mathbb{R}^n} \tag{15.24}$$

Note that $\lambda_{min}(A)$ and $\lambda_{max}(A)$ can be negative. The *lower bound* in (15.24) is met with equality by any eigenvector of A corresponding to the smallest eigenvalue; the *upper bound* in (15.24) is met with equality by any eigenvector of A corresponding to the largest eigenvalue. Similarly, from (15.23) it follows that

$$|x^\top A x| \le \sum_{i=1}^{n} c_i^2 |\lambda_i| \quad \forall x \in \mathbb{R}^n$$

and so

$$\boxed{|x^\top A x| \le \theta_{max}(A)\,||x||^2 \quad \forall x \in \mathbb{R}^n} \tag{15.25}$$

where equality holds for any eigenvector of A that corresponds to the eigenvalue that has largest absolute value.

Recall that a real matrix A is *symmetric and positive semidefinite* if $A^\top = A$ and

$$x^\top A x \ge 0 \quad \forall x \in \mathbb{R}^n$$

LEMMA 15.6. *Let A be a real and symmetric matrix.*

a) A is positive semidefinite if and only if all eigenvalues are nonnegative.

b) A is positive definite if and only if all eigenvalues are positive.

c) If B is any real $m \times n$ matrix then $B^\top B$ is symmetric positive semidefinite with all nonnegative eigenvalues.

PROOF. The proof follows directly from (15.23). □

15.6. Optimizing $x^\top B x$ and $|x^\top B x|$

Suppose B is a real $n \times n$ matrix (not necessarily symmetric). Observe that:

$$x^\top B x = \frac{1}{2} x^\top (B + B^\top) x \quad \forall x \in \mathbb{R}^n \tag{15.26}$$

This is true because $x^\top B x$ is a real number, which can be viewed as a 1×1 matrix, so it is equal to its transpose:

$$x^\top B x = (x^\top B x)^\top = x^\top B^\top x$$

This is significant because the matrix $\frac{1}{2}(B + B^\top)$ is real and symmetric! Thus, maximizing $x^\top Bx$ over a given constraint set is equivalent to maximizing $x^\top Ax$ over the desired set, where A is the real and symmetric matrix $A = \frac{1}{2}(B + B^\top)$.

Substituting $A = \frac{B+B^\top}{2}$ into (15.24) gives

$$\lambda_{min}\left(\frac{B + B^\top}{2}\right) \cdot ||x||^2 \leq x^\top \left(\frac{B + B^\top}{2}\right) x \leq \lambda_{max}\left(\frac{B + B^\top}{2}\right) \cdot ||x||^2 \quad \forall x \in \mathbb{R}^n$$

where the $\lambda_{min}(A)$, $\lambda_{max}(A)$ functions are defined for real and symmetric matrices A in (15.13) and (15.14). Using the identity (15.26) gives

$$\boxed{\lambda_{min}\left(\frac{B + B^\top}{2}\right) \cdot ||x||^2 \leq x^\top Bx \leq \lambda_{max}\left(\frac{B + B^\top}{2}\right) \cdot ||x||^2 \quad \forall x \in \mathbb{R}^n}$$

The lower bound is achieved by any eigenvector of $\frac{1}{2}(B+B^\top)$ that corresponds to eigenvalue $\lambda_{min}(\frac{B+B^\top}{2})$; the upper bound is achieved by any eigenvector of $\frac{1}{2}(B+B^\top)$ that corresponds to eigenvalue $\lambda_{max}(\frac{B+B^\top}{2})$.

Similarly, using (15.26) in (15.25) gives

$$\boxed{|x^\top Bx| \leq \theta_{max}\left(\frac{B + B^\top}{2}\right) \cdot ||x||^2 \quad \forall x \in \mathbb{R}^n}$$

where $\theta_{max}(A)$ is defined in (15.16). The upper bound is achieved by any eigenvector of $\frac{1}{2}(B + B^\top)$ that corresponds to the maximum magnitude eigenvalue.

Exercise 15.19 uses the equalities (15.26) and (15.23) to place a *lower bound* on $|x^\top Bx|$ in terms of $||x||^2$, which leads to the following interesting behavior: If $B + B^\top$ has at least one positive eigenvalue and at least one negative eigenvalue, the best lower bound on $|x^\top Bx|$ is zero! Otherwise, the best lower bound is $|x^\top Bx| \geq \theta_{min}(\frac{1}{2}(B + B^\top)) \cdot ||x||^2$ for all $x \in \mathbb{R}^n$.

15.7. Optimizing $||Bx||$

Let B be a real $m \times n$ matrix (not necessarily square). Note that

$$\begin{aligned} ||Bx||^2 &= (Bx)^\top (Bx) \\ &= x^\top B^\top Bx \\ &= |x^\top B^\top Bx| \end{aligned}$$

Therefore, minimizing or maximizing $||Bx||$ over a given constraint set is equivalent to minimizing or maximizing $|x^\top Ax|$, where A is the real and symmetric matrix $A = B^\top B$. Substituting $A = B^\top B$ into (15.24) gives

$$\lambda_{min}\left(B^\top B\right)||x||^2 \leq \underbrace{\left|x^\top B^\top Bx\right|}_{||Bx||^2} \leq \lambda_{max}\left(B^\top B\right)||x||^2 \quad \forall x \in \mathbb{R}^n$$

where the $\lambda_{min}(A)$ and $\lambda_{max}(A)$ functions are defined in (15.13) and (15.14). Note that $B^\top B$ is positive semidefinite with nonnegative eigenvalues (Lemma 15.6c). Taking square roots gives (for any real $m \times n$ matrix B):

$$\boxed{\sqrt{\lambda_{min}\left(B^\top B\right)} \cdot ||x|| \leq ||Bx|| \leq \sqrt{\lambda_{max}\left(B^\top B\right)} \cdot ||x|| \quad \forall x \in \mathbb{R}^n}$$

The upper bound is achieved by any eigenvector of $B^\top B$ that corresponds to maximum eigenvalue; the lower bound corresponds to any eigenvector of $B^\top B$ that corresponds to the minimum eigenvalue. In Chapter 16 we learn that $\sqrt{\lambda_{max}(B^\top B)}$ is the *largest singular value* of matrix B, while $\sqrt{\lambda_{min}(B^\top B)}$ is the *smallest singular value* of matrix B.

15.8. Example with three distinct eigenvalues

Consider the real and symmetric matrix

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

Clearly -2 is an eigenvalue with eigenvector $(1, 0, 0)$, so the matrix is not positive semidefinite. However, our theory says that the remaining eigenvalues must be real and there must be three eigenvectors that form an orthonormal basis for $\mathbb{R}^3$. Eigenvalues can be found by solving $\det(A - \lambda I) = 0$. Since this matrix has a block triangular structure we have

$$0 = \det(A - \lambda I) = (-2 - \lambda)\left[(3 - \lambda)(4 - \lambda) - 1\right] = (-2 - \lambda)(\lambda^2 - 7\lambda + 11)$$

$$\begin{aligned} \lambda_1 &= -2 \\ \lambda_2 &= \frac{7 + \sqrt{5}}{2} = 4.6180 \\ \lambda_3 &= \frac{7 - \sqrt{5}}{2} = 2.3820 \end{aligned}$$

We see that all eigenvalues are indeed real. The three eigenvalues are distinct, so *any* eigenvectors we find for them will automatically be orthogonal to each other. We already know an eigenvector for λ_1 is $(1,0,0)$. To find nonzero vectors in $Null(A-\lambda I)$ for $\lambda \in \{\lambda_2, \lambda_3\}$ we use Gaussian elimination:

$$\begin{bmatrix} -2-\lambda & 0 & 0 \\ 0 & 3-\lambda & 1 \\ 0 & 1 & 4-\lambda \end{bmatrix} \to \begin{bmatrix} -2-\lambda & 0 & 0 \\ 0 & 1 & 4-\lambda \\ 0 & 3-\lambda & 1 \end{bmatrix} \to \begin{bmatrix} -2-\lambda & 0 & 0 \\ 0 & 1 & 4-\lambda \\ 0 & 0 & 0 \end{bmatrix}$$

where the last row of zeros follows because we know the matrix is not invertible (and hence cannot have 3 pivots) for $\lambda \in \{\lambda_2, \lambda_3\}$. Considering eigenvectors of the form $z = (0, a, 1)$, we have

$$a + (4-\lambda) = 0 \implies a = \lambda - 4$$

Thus

$$\begin{aligned} \lambda_1 &= -2, \quad z_1 = (1,0,0) \\ \lambda_2 &= \frac{7+\sqrt{5}}{2}, \quad z_2 = \left(0, \frac{-1+\sqrt{5}}{2}, 1\right) \\ \lambda_3 &= \frac{7-\sqrt{5}}{2}, \quad z_3 = \left(0, \frac{-1-\sqrt{5}}{2}, 1\right) \end{aligned}$$

The eigenvectors $\{z_1, z_2, z_3\}$ are indeed orthogonal. It remains to normalize them:

$$\begin{aligned} v_1 &= (1,0,0) \\ v_2 &= z_2 / ||z_2|| = (0, 0.5257, 0.8507) \\ v_3 &= z_3 / ||z_3|| = (0, -0.8507, 0.5257) \end{aligned}$$

and we indeed have

$$A = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5257 & -0.8507 \\ 0 & 0.8507 & 0.5257 \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} -2 & 0 & 0 \\ 0 & 4.6180 & 0 \\ 0 & 0 & 2.3820 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5257 & -0.8507 \\ 0 & 0.8507 & 0.5257 \end{bmatrix}^{\top}}_{S^{\top}}$$

and $S^{\top}S = I$, so $S^{\top} = S^{-1}$. Since the minimum absolute value of all eigenvalues is 2, and the maximum absolute value is 4.6180, we have

$$2\,||x|| \le ||Ax|| \le 4.6180\,||x|| \quad \forall x \in \mathbb{R}^3$$

where the lower bound is achieved by any vector $x \in \mathbb{R}^3$ that is a scalar multiple of v_1, and the upper bound is achieved by any $x \in \mathbb{R}^3$ that is a scalar multiple of v_2.

15.9. Example with a repeated eigenvalue

Consider the following real and symmetric matrix:

$$A = \begin{bmatrix} 11 & -1 & -\sqrt{2} \\ -1 & 11 & -\sqrt{2} \\ -\sqrt{2} & -\sqrt{2} & 10 \end{bmatrix}$$

This matrix has been crafted to have only two distinct eigenvalues, so some care must be taken to find corresponding orthonormal eigenvectors for the repeated eigenvalue. To find the eigenvalues we find λ to ensure $Null(A - \lambda I)$ is a nontrivial subspace. Using Gaussian elimination gives

$$\begin{aligned} A - \lambda I = \begin{bmatrix} 11-\lambda & -1 & -\sqrt{2} \\ -1 & 11-\lambda & -\sqrt{2} \\ -\sqrt{2} & -\sqrt{2} & 10-\lambda \end{bmatrix} &\overset{(a)}{\to} \begin{bmatrix} -1 & 11-\lambda & -\sqrt{2} \\ -\sqrt{2} & -\sqrt{2} & 10-\lambda \\ 11-\lambda & -1 & -\sqrt{2} \end{bmatrix} \\ &\overset{(b)}{\to} \begin{bmatrix} -1 & 11-\lambda & -\sqrt{2} \\ 0 & -\sqrt{2}(12-\lambda) & 12-\lambda \\ 11-\lambda & -1 & -\sqrt{2} \end{bmatrix} \\ &\overset{(c)}{\to} \begin{bmatrix} -1 & 11-\lambda & -\sqrt{2} \\ 0 & -\sqrt{2}(12-\lambda) & 12-\lambda \\ 0 & (11-\lambda)^2 - 1 & -\sqrt{2}(12-\lambda) \end{bmatrix} \end{aligned} \tag{15.27}$$

where step (a) permutes the rows to place a pivot in the upper left corner and a $-\sqrt{2}$ underneath it; step (b) multiplies the first row by $-\sqrt{2}$ and adds it into the second row; step (c) multiplies the first row by $11 - \lambda$ and adds it into the last row. One way to make the resulting matrix not invertible is to choose $\lambda = 12$ (which is motivated by making the middle diagonal entry in the right-hand-side of (15.27) a 0 rather than a pivot):

$$\lambda = 12 \implies A - \lambda I \to \begin{bmatrix} -1 & -1 & -\sqrt{2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

which means we can find two orthogonal vectors in this null space, so we have

$$\boxed{\lambda_1 = \lambda_2 = 12}$$

and we seek a basis of eigenvectors for this 2-dimensional null space using free variables b, c:

$$\begin{bmatrix} -1 & -1 & -\sqrt{2} \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = 0 \iff \begin{bmatrix} a \\ b \\ c \end{bmatrix} = b \begin{bmatrix} -1 \\ 1 \\ 0 \end{bmatrix} + c \begin{bmatrix} -\sqrt{2} \\ 0 \\ 1 \end{bmatrix}$$

The vectors $w_1 = (-1, 1, 0)$ and $w_2 = (-\sqrt{2}, 0, 1)$ are indeed eigenvectors of A with eigenvalue 12, but they are not orthogonal. Since w_1 and w_2 have the same eigenvalue 12, any nonzero linear combination of w_1 and w_2 is also an eigenvector. We take z_1 as the first vector and we take z_2 as the projection of the second onto the space orthogonal to the first:

$$\begin{aligned} z_1 &= w_1 = (-1, 1, 0) \\ z_2 &= w_2 - \frac{(w_2^\top w_1) w_1}{\|w_1\|^2} = \left(-\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 1 \right) \end{aligned}$$

It can be verified that z_1, z_2 are orthogonal and they remain eigenvectors of A with eigenvalue 12. It remains to normalize them:

$$\boxed{v_1 = z_1 / \|z_1\| = \left(\frac{-1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0 \right)}$$

$$\boxed{v_2 = z_2 / \|z_2\| = \left(\frac{-1}{2}, \frac{-1}{2}, \frac{1}{\sqrt{2}} \right)}$$

and one can verify $Av_1 = 12v_1$, $Av_2 = 12v_2$, $\|v_1\| = \|v_2\| = 1$, $v_1^\top v_2 = 0$.

We need to find the remaining eigenvalue $\lambda_3 \neq 12$. Fortunately, the eigenvector associated with this eigenvalue will automatically be orthogonal to both v_1 and v_2. Assuming $\lambda \neq 12$ in the right-hand-side of (15.27) makes the middle diagonal entry a pivot (nonzero) and so

$$A - \lambda I \overset{(a)}{\to} \begin{bmatrix} -1 & 11 - \lambda & -\sqrt{2} \\ 0 & -\sqrt{2}(12 - \lambda) & 12 - \lambda \\ 0 & (11 - \lambda)^2 - 1 & -\sqrt{2}(12 - \lambda) \end{bmatrix} \overset{(b)}{\to} \begin{bmatrix} -1 & 11 - \lambda & -\sqrt{2} \\ 0 & 1 & -1/\sqrt{2} \\ 0 & (11 - \lambda)^2 - 1 & -\sqrt{2}(12 - \lambda) \end{bmatrix}$$

where step (a) simply copies the result of the right-hand-side of (15.27); step (b) divides row 2 by the nonzero value $-\sqrt{2}(12 - \lambda)$. To make the matrix not invertible, we want the

last row to be a perfect multiple of the second row:

$$-\sqrt{2}(12-\lambda) = (-1/\sqrt{2})\left[(11-\lambda)^2 - 1\right]$$

Solving this quadratic equation gives a root of 12 (an eigenvalue we already have) and a new root of 8, so

$$\boxed{\lambda_3 = 8}$$

With this eigenvalue $\lambda = 8$ we continue the Gaussian elimination:

$$A - \lambda I \to \begin{bmatrix} -1 & 3 & -\sqrt{2} \\ 0 & 1 & -1/\sqrt{2} \\ 0 & 8 & -4\sqrt{2} \end{bmatrix} \to \begin{bmatrix} -1 & 3 & -\sqrt{2} \\ 0 & 1 & -1/\sqrt{2} \\ 0 & 0 & 0 \end{bmatrix}$$

The corresponding eigenvector $z_3 = (a, b, 1)$ solves

$$\begin{aligned} -a + 3b - \sqrt{2} &= 0 \\ b - 1/\sqrt{2} &= 0 \end{aligned}$$

So

$$a = b = 1/\sqrt{2} \implies z_3 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 1\right)$$

$$\boxed{v_3 = \frac{z_3}{||z_3||} = \left(\frac{1}{2}, \frac{1}{2}, \frac{1}{\sqrt{2}}\right)}$$

One can verify $Av_3 = 8v_3$, $||v_3|| = 1$, and $v_3^\top v_1 = 0$, $v_3^\top v_2 = 0$. Overall we have

$$A = \begin{bmatrix} 11 & -1 & -\sqrt{2} \\ -1 & 11 & -\sqrt{2} \\ -\sqrt{2} & -\sqrt{2} & 10 \end{bmatrix} = \underbrace{\begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{2} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{-1}{2} & \frac{1}{2} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} 12 & 0 & 0 \\ 0 & 12 & 0 \\ 0 & 0 & 8 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{-1}{2} & \frac{1}{2} \\ \frac{1}{\sqrt{2}} & \frac{-1}{2} & \frac{1}{2} \\ 0 & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}^\top}_{S^\top}$$

and $S^\top S = I$, so $S^\top = S^{-1}$. Since all eigenvalues of this real and symmetric matrix A are positive, we know A is positive definite. Also, since the lowest magnitude eigenvalue of A is 8 and the largest magnitude eigenvalue is 12, we have

$$8\,||x|| \le ||Ax|| \le 12\,||x|| \qquad \forall x \in \mathbb{R}^3$$

where the lower bound is achieved by any scalar multiple of v_3 and the upper bound is

achieved by any linear combination of v_1, v_2.

15.10. Example maximization of $|x^\top Bx|$

Consider the problem of finding a vector $x \in \mathbb{R}^4$ to solve

$$\text{Maximize:} \quad |x^\top Bx| \tag{15.28}$$

$$\text{Subject to:} \quad ||x|| = 77 \tag{15.29}$$

$$x \in \mathbb{R}^4 \tag{15.30}$$

where $||x|| = \sqrt{x^\top x}$ and B is the real (but not symmetric) matrix

$$B = \begin{bmatrix} 5 & -2 & 0 & 8 \\ 3 & 8 & 0 & 1 \\ -7 & 2 & 0 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

We call this "problem (15.28)-(15.30)," which refers to the problem defined by (15.28), (15.29), and (15.30).

Define the real and symmetric matrix A by

$$A = \frac{1}{2}(B + B^\top) = \begin{bmatrix} 5 & 0.5 & -3.5 & 4.5 \\ 0.5 & 8 & 1 & 1 \\ -3.5 & 1 & 0 & 0.5 \\ 4.5 & 1 & 0.5 & 1 \end{bmatrix}$$

We know from (15.26) that $x^\top Bx = x^\top Ax$ for all $x \in \mathbb{R}^4$. From results of Section 15.6 we know

$$|x^\top Bx| \le \theta_{max}(A)\, ||x||^2 \quad \forall x \in \mathbb{R}^4$$

where $\theta_{max}(A)$ is the magnitude of the largest-magnitude eigenvalue of A, and equality is achieved by any corresponding eigenvector. In particular

$$||x|| = 77 \implies |x^\top Bx| \le 77^2 \theta_{max}(A) \tag{15.31}$$

To solve problem (15.28)-(15.30), we need to find the (real) orthonormal eigenvectors $\{v_1, ..., v_4\}$ of A and the corresponding (real) eigenvalues $\lambda_1, ..., \lambda_4$. We use the Matlab command:

$$[\texttt{S}, \texttt{Lambda}] = \texttt{eig}(\texttt{A})$$

which gives

$$S = \begin{bmatrix} -0.5454 & 0.0907 & -0.4518 & -0.7001 \\ 0.0251 & 0.1821 & 0.8278 & -0.5301 \\ -0.6027 & -0.7176 & 0.2953 & 0.1860 \\ \underbrace{0.5820}_{v_1} & \underbrace{-0.6660}_{v_2} & \underbrace{-0.1533}_{v_3} & \underbrace{-0.4407}_{v_4} \end{bmatrix}, \quad \Lambda = \begin{bmatrix} -3.6920 & 0 & 0 & 0 \\ 0 & 0.6525 & 0 & 0 \\ 0 & 0 & 7.8986 & 0 \\ 0 & 0 & 0 & 9.1409 \end{bmatrix}$$

It can be verified that $||v_i|| = 1$ and $v_i^\top v_j = 0$ for all $i \neq j$. From this information we obtain

$$\begin{aligned} |\lambda_1| &= 3.6920 \\ |\lambda_2| &= 0.6525 = \theta_{min}(A) \\ |\lambda_3| &= 7.8986 \\ |\lambda_4| &= 9.1409 = \theta_{max}(A) \end{aligned}$$

and so from (15.31)

$$||x|| = 77 \implies |x^\top Bx| \leq 77^2(9.1409) = 54196.1973$$

with equality when $x = 77v_4$. Therefore, the optimal $x^* \in \mathbb{R}^4$ to solve problem (15.28)-(15.30) is

$$x^* = 77v_4 = \begin{bmatrix} -53.9096 \\ -40.8187 \\ 14.3201 \\ -33.9340 \end{bmatrix} \tag{15.32}$$

and the optimal objective value is $|(x^*)^\top Bx^*| = 54196.1973$. Of course, the vector $-x^*$ is equally optimal.[1]

[1]With slightly more work, it can be shown that problem (15.28)-(15.30) has *exactly two optimal solutions*, being x^* and $-x^*$ (where x^* is given in (15.32)). Readers familiar with the definitions of *convex set* and *convex function* may be interested to note that the nonconvex constraint $||x|| = 77$ can be changed to the convex constraint $||x|| \leq 77$ without changing the solution to the optimization problem (15.28)-(15.30). This is because any nonzero vector that satisfies $||x|| < 77$ can be scaled to have norm exactly 77, which increases the value of $|x^\top Bx|$. However, the problem remains a nonconvex optimization problem under this change because maximizing $|x^\top Bx|$ is equivalent to minimizing $-|x^\top Bx|$, but the function $-|x^\top Bx|$ is nonconvex in x. Nonconvexity of the problem (15.28)-(15.30) is also apparent from the fact that it is impossible for a convex optimization problem to have exactly two optimal solutions. This is because convex optimization problems always have convex solution sets, so if there are two distinct solutions then all (infinitely many) points on the line segment between those solutions must also be a solution.

Our particular 4×4 matrix B is invertible and has one negative eigenvalue and three positive eigenvalues. Exercise 15.19 shows that if problem (15.28)-(15.30) were changed to *minimizing* $|x^\top Bx|$ subject to the same constraints, then (even though $Null(B)$ is trivial) the answer is a norm-77 vector $\tilde{x}^* \in \mathbb{R}^4$ that has an optimal objective value of zero!

15.11. Skew symmetric and skew Hermitian matrices

- A square matrix A is *skew symmetric* if $A^\top = -A$.
- A square matrix A with complex entries is *skew Hermitian* if $A^H = -A$.

It is easy to see that real skew symmetric matrices are also skew Hermitian, and all skew Hermitian matrices are normal. Exercise 15.9 shows that all skew Hermitian matrices have purely imaginary eigenvalues. Examples are

- Real and skew symmetric:

$$A = \begin{bmatrix} 0 & 1 & 2 \\ -1 & 0 & 3 \\ -2 & -3 & 0 \end{bmatrix} = S \begin{bmatrix} 0 & 0 & 0 \\ 0 & 3.7417i & 0 \\ 0 & 0 & -3.7417i \end{bmatrix} S^H$$

  where

$$S = \begin{bmatrix} 0.8018 & 0.2988 + 0.2988i & 0.2988 - 0.2988i \\ -0.5345 & 0.1350 + 0.5822i & 0.1350 - 0.5822i \\ 0.2673 & -0.6265 + 0.2679i & -0.6265 - 0.2679i \end{bmatrix}$$

  and it holds that $S^H S = I$. The columns of S are eigenvectors of A and form an orthonormal basis for $\mathbb{C}^3$.

- Skew Hermitian:

$$A = \begin{bmatrix} 0 & 1 & 0 \\ -1 & 0 & 7 + i \\ 0 & -7 + i & i \end{bmatrix} = S \begin{bmatrix} 7.6498i & 0 & 0 \\ 0 & 0.0196i & 0 \\ 0 & 0 & -6.6694i \end{bmatrix} S^H$$

  where

$$S = \begin{bmatrix} 0.0883 + 0.0126i & -0.9800 - 0.1400i & -0.1085 - 0.0155i \\ -0.0965 + 0.6755i & 0.0027 - 0.0192i & -0.1033 + 0.7234i \\ -0.7256 + 0.0000i & -0.1399 + 0.0000i & 0.6738 + 0.0000i \end{bmatrix}$$

  and it holds that $S^H S = I$. The columns of S are eigenvectors of A and form an orthonormal basis for $\mathbb{C}^3$.

15.12. Additional properties of normal matrices

LEMMA 15.7. *Let A be a real and symmetric $n \times n$ matrix that satisfies*

$$x^\top A x = 0 \quad \forall x \in \mathbb{R}^n \tag{15.33}$$

Then $A = 0$ (the all zero matrix).

PROOF. We know that all real and symmetric matrices have an orthonormal basis of real eigenvectors $\{v_1, ..., v_n\}$ with real eigenvalues $\lambda_1, ..., \lambda_n$, and

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^\top$$

However, it is easy to see that (15.33) implies all eigenvalues of A are 0 (just substitute $x = v_i$ into (15.33)). So $A = 0$. □

LEMMA 15.8. *Let A be* any *$n \times n$ matrix with complex entries that satisfies*

$$x^H A x = 0 \quad \forall x \in \mathbb{C}^n \tag{15.34}$$

Then $A = 0$ (the all zero matrix).

PROOF. We prove the result by first making a weaker claim:

Claim: If A is a *normal* $n \times n$ matrix that satisfies (15.34), then $A = 0$.

To prove the claim, we know that every $n \times n$ normal matrix A has an orthonormal basis of eigenvectors $\{v_1, ..., v_n\}$ with eigenvalues $\lambda_1, ..., \lambda_n$ and

$$A = \sum_{i=1}^{n} \lambda_i v_i v_i^H$$

On the other hand, it is easy to see that all eigenvalues of A are zero (just substitute $x = v_i$ into (15.34)). So $A = 0$.

Now suppose A is *any* $n \times n$ matrix with complex entries that satisfies (15.34) (not necessarily normal). Taking Hermitians of (15.34) yields

$$x^H A^H x = 0 \quad \forall x \in \mathbb{C}^n$$

and this together with condition (15.34) implies:

$$x^H(A + A^H)x = 0 \quad \forall x \in \mathbb{C}^n \tag{15.35}$$

$$x^H(A - A^H)x = 0 \quad \forall x \in \mathbb{C}^n \tag{15.36}$$

Since $A + A^H$ is Hermitian, it is normal, and the condition (15.35) together with our claim implies

$$A + A^H = 0 \tag{15.37}$$

Likewise, since $A - A^H$ is skew Hermitian, it is normal, and the condition (15.36) together with our claim implies

$$A - A^H = 0 \tag{15.38}$$

Adding (15.37) and (15.38) gives $2A = 0$, so $A = 0$. □

It is interesting that Lemma 15.7 requires A to be real and symmetric, while Lemma 15.8 places no restrictions on the square matrix A. This is because the condition (15.34) is more stringent and must hold for all $x \in \mathbb{C}^n$, whereas the condition (15.33) only holds for all $x \in \mathbb{R}^n$. For example, consider the fact

$$\begin{bmatrix} x & y \end{bmatrix} \underbrace{\begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix}}_{A} \begin{bmatrix} x \\ y \end{bmatrix} = 0 \quad \forall (x, y) \in \mathbb{R}^2 \tag{15.39}$$

which is true because the equation reduces to $xy - xy = 0$ for all $(x, y) \in \mathbb{R}^2$. Of course, this matrix A is not the zero matrix. This does not contradict Lemma 15.7 because, even though A is real, it is not symmetric. This example also does not contradict Lemma 15.8 because the required condition for that lemma fails to hold for all $x \in \mathbb{C}^n$. Indeed, we observe

$$\begin{bmatrix} 1 \\ i \end{bmatrix}^H \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ i \end{bmatrix} = \begin{bmatrix} 1 & -i \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} \begin{bmatrix} 1 \\ i \end{bmatrix} = 2i \neq 0$$

Lemma 15.9. *A $n \times n$ matrix A with complex entries is normal if and only if*

$$||Ax|| = \left|\left|A^H x\right|\right| \quad \forall x \in \mathbb{C}^n \tag{15.40}$$

PROOF. For the forward statement, first suppose A is normal. The definition of normal directly implies that (15.40) holds (see Exercise 15.8). For the reverse statement, suppose (15.40) holds. We want to show A is normal. Fix $x \in \mathbb{C}^n$. We have

$$\begin{aligned} 0 &= \|Ax\|^2 - \left\|A^H x\right\|^2 \\ &= (Ax)^H (Ax) - (A^H x)^H (A^H x) \\ &= x^H A^H A x - x^H A A^H x \\ &= x^H (A^H A - A A^H) x \end{aligned}$$

This holds for all $x \in \mathbb{C}^n$, and so

$$x^H (A^H A - A A^H) x = 0 \quad \forall x \in \mathbb{C}^n$$

so Lemma 15.8 implies $A^H A - A A^H = 0$, meaning that A is normal. □

15.13. Exercises

EXERCISE 15.1. *Consider the real and symmetric matrix*

$$A = \begin{bmatrix} 3 & 10 \\ 10 & 3 \end{bmatrix}$$

a) Find the eigenvalues for A. Are they real?

b) Find corresponding eigenvectors and show they are orthogonal.

c) Diagonalize A using an orthonormal basis of eigenvectors.

EXERCISE 15.2. *Consider the real and symmetric matrix*

$$A = \begin{bmatrix} -2 & 1 \\ 1 & 4 \end{bmatrix}$$

a) Find the eigenvalues for A. Are they real?

b) Find corresponding eigenvectors and show they are orthogonal.

c) Diagonalize A using an orthonormal basis of eigenvectors.

EXERCISE 15.3. *Consider the real and symmetric matrix*

$$A = \begin{bmatrix} 2 & 0 & 1 \\ 0 & 3 & 9 \\ 1 & 9 & 4 \end{bmatrix}$$

a) Find the three eigenvalues for A. Are they real?

b) Find corresponding three eigenvectors and show they are orthogonal.

c) Diagonalize A using an orthonormal basis of eigenvectors.

EXERCISE 15.4. *(A symmetric and positive definite matrix) Consider the real and symmetric matrix*

$$A = \begin{bmatrix} 5 & 1 \\ 1 & 5 \end{bmatrix}$$

a) Find the eigenvalues for A. Are they real? Are they positive?

b) Find corresponding eigenvectors and show they are orthogonal.

c) Diagonalize A using an orthonormal basis of eigenvectors.

d) Show that $x^\top Ax \geq 0$ for all $x \in \mathbb{R}^2$, with equality if and only if $x = 0$. Thus, the matrix is symmetric and positive definite.

EXERCISE 15.5. *(Real symmetric 2×2 matrices) Consider a real 2×2 symmetric matrix*

$$A = \begin{bmatrix} a & b \\ b & d \end{bmatrix}$$

for some constants $a, b, d \in \mathbb{R}$.

a) Compute the eigenvalues λ_1, λ_2 in terms of a, b, d using the quadratic formula on the equation $\det(A - \lambda I) = 0$. Show that, using your expressions for λ_1 and λ_2, both eigenvalues are real. This involves showing that an expression inside a square root is always nonnegative for all $a, b, d \in \mathbb{R}$.

b) Choose any specific numerical values for $a, b, d \in \mathbb{R}$ under which $\lambda_1 \neq \lambda_2$. Numerically compute the corresponding eigenvectors $v_1, v_2 \in \mathbb{R}^2$ and show $v_1^\top v_2 = 0$.

EXERCISE 15.6. *Suppose A is a $n \times n$ Hermitian matrix, so A has complex entries and $A^H = A$. Fix k as an integer such that $1 \leq k < n$. Suppose $\{v_1, \ldots, v_k\}$ are eigenvectors of A. Define*

$$S = Span\{v_1, \ldots, v_k\}^\perp$$

where we use the standard inner product $\langle x, y \rangle = x^H y$ for all $x, y \in \mathbb{C}^n$. Show that S is an invariant subspace under the linear map A. That is, fix $x \in S$ and show $Ax \in S$.

EXERCISE 15.7. *(Distinct eigenvalues of a real and symmetric matrix have orthogonal eigenvectors) Let A be a real and symmetric $n \times n$ matrix. We know that all eigenvalues of A are real (Lemma 15.3), so their eigenvectors can be chosen to be real (Lemma 15.1). Let (λ_1, v_1) and (λ_2, v_2) be two eigenpairs of A such that $\lambda_1, \lambda_2 \in \mathbb{R}$ and $v_1, v_2 \in \mathbb{R}^n$. Assume $\lambda_1 \neq \lambda_2$. Use the standard inner product $\langle x, y \rangle = x^\top y$ for all $x, y \in \mathbb{R}^n$.*

a) Starting with the equation $\lambda_1 v_1 = Av_1$, take transposes of both sides to find an expression

for $\lambda_1 v_1^\top$.

b) Multiply the equation of (a) on the right by v_2 to obtain an expression for $\lambda_1 \langle v_1, v_2 \rangle$.

c) Use part (b) (and the facts $A^\top = A$ and $\lambda_1 \neq \lambda_2$) to show that v_1 and v_2 are orthogonal.

d) Give a specific numerical example of a real and symmetric 2×2 matrix $A = (a_{ij})$ with $a_{12} = a_{21} \neq 0$. Numerically compute the eigenvalues and show they are distinct (so $\lambda_1 \neq \lambda_2$). Compute the corresponding eigenvectors $v_1, v_2 \in \mathbb{R}^2$ and show they are orthogonal.

EXERCISE 15.8. *(Normal $\implies ||Ax|| = \left||A^H x\right||$) Let A be a $n \times n$ matrix with complex entries. Suppose A is normal, so $A^H A = AA^H$. Prove that for all $x \in \mathbb{C}^n$ we have*

$$||Ax||^2 = \left||A^H x\right||^2$$

where the norm is defined $||z||^2 = z^H z$ for all $z \in \mathbb{C}^n$.

EXERCISE 15.9. *(Skew Hermitian has imaginary eigenvalues) Let A be a square matrix with complex entries. Suppose A is skew Hermitian, so $A^H = -A$.*

a) Show that if λ is an eigenvalue of A, then $\overline{\lambda} = -\lambda$. Hint: Use Lemma 15.4.

b) Conclude that all eigenvalues of A are purely imaginary (meaning a real multiple of i).

c) Show that all diagonal entries of a skew Hermitian matrix are purely imaginary.

d) Compute the eigenvalues of the following skew Hermitian matrices:

$$A = \begin{bmatrix} 0 & 8 \\ -8 & 0 \end{bmatrix}$$

$$B = \begin{bmatrix} i & 2 \\ -2 & i \end{bmatrix}$$

EXERCISE 15.10. *(Maximizing $|x^\top Ax|$) Solve the problem*

$$\begin{aligned} \textit{Maximize:} \quad & |x^\top Ax| \\ \textit{Subject to:} \quad & ||x|| = 3 \\ & x \in \mathbb{R}^2 \end{aligned}$$

for

$$A = \begin{bmatrix} -5 & 1 \\ 1 & 2 \end{bmatrix}$$

EXERCISE 15.11. *(Maximizing $x^\top Bx$ and $||Bx||$) Let B be a real $n \times n$ matrix (not necessarily symmetric). Let $||x|| = \sqrt{x^\top x}$ denote the standard Euclidean norm. Define the following*

nonnegative real numbers:

$$\sigma = \max_{\{x \in \mathbb{R}^n : ||x||=1\}} \{||Bx||\}$$
$$\theta = \max_{\{x \in \mathbb{R}^n : ||x||=1\}} \left\{|x^\top Bx|\right\}$$

We want to show that $\theta \leq \sigma$*, and if* B *is real and symmetric then* $\theta = \sigma$*.*

a) Use the Cauchy-Schwarz inequality to show $\theta \leq \sigma$*. Hint:* $x^\top Bx = \langle x, Bx \rangle$*.*

b) Suppose B *is real and symmetric with (necessarily real) eigenvalues* $\lambda_1, \ldots, \lambda_n$*. Show that*

$$\theta = \sigma = \max_{i \in \{1,\ldots,n\}} \{|\lambda_i|\}$$

c) Find a vector $x^* \in \mathbb{R}^2$ *that maximizes* $|x^\top Bx|$ *subject to* $||x|| = 8.7$ *when* B *is the real and symmetric matrix:*

$$B = \begin{bmatrix} 2 & 3 \\ 3 & -5 \end{bmatrix}$$

EXERCISE 15.12. *(*$\theta_{min}(A)$ *and* $\theta_{max}(A)$*) As in Section 15.6, for real and symmetric matrices* A *define* $\theta_{min}(A)$ *and* $\theta_{max}(A)$ *as the magnitude of the min-magnitude eigenvalue of* A*, and max-magnitude eigenvalue of* A*, respectively.*

a) Show that if A *is a real and symmetric matrix then for all* $c \in \mathbb{R}$ *we have*

$$\theta_{min}(cA) = |c|\theta_{min}(A)$$
$$\theta_{max}(cA) = |c|\theta_{max}(A)$$

b) Compute $\theta_{min}(A)$*,* $\theta_{max}(A)$*,* $\theta_{min}(A/2)$*,* $\theta_{max}(A/2)$*,* $\theta_{min}(-A/2)$*,* $\theta_{max}(-A/2)$ *for*

$$A = \begin{bmatrix} 3 & 9 \\ 9 & -8 \end{bmatrix}$$

EXERCISE 15.13. *(Maximizing and minimizing* $x^\top Ax$*) Consider*

$$A = \begin{bmatrix} 10 & 10 \\ 10 & 10 \end{bmatrix}$$

a) Solve the problem

$$\text{Maximize:} \quad x^\top Ax$$
$$\text{Subject to:} \quad ||x|| = 11$$

$$x \in \mathbb{R}^2$$

b) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Ax \\ \textit{Subject to:} \quad & ||x|| = 11 \\ & x \in \mathbb{R}^2 \end{aligned}$$

EXERCISE 15.14. *(More maximizing and minimizing $x^\top Ax$) Consider*

$$A = \begin{bmatrix} 10 & -1 \\ -1 & 10 \end{bmatrix}$$

a) Solve the problem

$$\begin{aligned} \textit{Maximize:} \quad & x^\top Ax \\ \textit{Subject to:} \quad & ||x|| = 5.1 \\ & x \in \mathbb{R}^2 \end{aligned}$$

b) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Ax \\ \textit{Subject to:} \quad & ||x|| = 5.1 \\ & x \in \mathbb{R}^2 \end{aligned}$$

EXERCISE 15.15. *(Using a computer to minimize and maximize $x^\top Ax$) Consider the real and symmetric matrix A given by*

$$A = \begin{bmatrix} -2 & 0 & 2 & 0 & 9 \\ 0 & -10 & 8 & 1 & 3 \\ 2 & 8 & -8 & 9 & 0 \\ 0 & 1 & 9 & 1 & 0 \\ 9 & 3 & 0 & 0 & 2 \end{bmatrix}$$

For this problem, you are encouraged to use a computer to compute the eigenvalues/vectors for A.

a) Solve the problem

$$\textit{Maximize:} \quad x^\top Ax$$

$$\begin{aligned} \textit{Subject to:} \quad & ||x|| = 7 \\ & x \in \mathbb{R}^5 \end{aligned}$$

b) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Ax \\ \textit{Subject to:} \quad & ||x|| = 7 \\ & x \in \mathbb{R}^5 \end{aligned}$$

c) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Ax \\ \textit{Subject to:} \quad & ||x|| = 21 \\ & x \in \mathbb{R}^5 \end{aligned}$$

EXERCISE 15.16. *(Using a computer to minimize and maximize $x^\top Bx$) Consider the real matrix B given by*

$$B = \begin{bmatrix} 8 & 2 & 0 & 0 & 3 \\ 1 & 1 & 1 & 2 & 2 \\ 33 & -2 & 9 & 0 & 1 \\ 3 & 0 & 0 & 1 & 0 \\ 2 & -2 & 2 & -1 & 1 \end{bmatrix}$$

For this problem, you are encouraged to use a computer to compute the eigenvalues/vectors for $\frac{1}{2}(B + B^\top)$.

a) Solve the problem

$$\begin{aligned} \textit{Maximize:} \quad & x^\top Bx \\ \textit{Subject to:} \quad & ||x|| = 7 \\ & x \in \mathbb{R}^5 \end{aligned}$$

b) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Bx \\ \textit{Subject to:} \quad & ||x|| = 7 \\ & x \in \mathbb{R}^5 \end{aligned}$$

c) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Bx \\ \textit{Subject to:} \quad & ||x|| = 21 \\ & x \in \mathbb{R}^5 \end{aligned}$$

EXERCISE 15.17. *(Using a computer to minimize and maximize $x^\top Bx$) Consider the real matrix B given by*

$$B = \begin{bmatrix} 4 & 3 & -1 & 7 & 1 & 1 & 0 \\ 0 & -3 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & 0 & 0 \\ 12 & 0 & 0 & 1 & 0 & 0 & 2 \\ 4 & 4 & 4 & 4 & 4 & 4 & 4 \\ 2 & -7 & -8 & 1 & 0 & 1 & 0 \\ -1 & -1 & -2 & -2 & -3 & -3 & -1 \end{bmatrix}$$

For this problem, you are encouraged to use a computer to compute the eigenvalues/vectors for $\frac{1}{2}(B + B^\top)$.

a) Solve the problem

$$\begin{aligned} \textit{Maximize:} \quad & x^\top Bx \\ \textit{Subject to:} \quad & ||x|| = 100 \\ & x \in \mathbb{R}^7 \end{aligned}$$

b) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Bx \\ \textit{Subject to:} \quad & ||x|| = 100 \\ & x \in \mathbb{R}^7 \end{aligned}$$

c) Solve the problem

$$\begin{aligned} \textit{Minimize:} \quad & x^\top Bx \\ \textit{Subject to:} \quad & ||x|| = 2 \\ & x \in \mathbb{R}^7 \end{aligned}$$

EXERCISE 15.18. *(Again using a computer to minimize and maximize $x^\top Bx$) Consider the real*

matrix B given by

$$B = \begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 2 & 2 & 8 & -8 & -10 & -1 & 1 & 0 & 0 & -1 \\ 3 & 3 & 0 & 1 & 0 & 1 & 0 & 0 & 1 & 1 \\ 0 & 0 & 7 & 0 & 1 & 1 & 0 & 0 & 2 & 2 \\ 1 & 1 & -1 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 2 & 2 & 5 & 5 & 5 & 5 & 5 & 5 & 5 & 5 \\ 3 & 3 & 8 & 8 & 8 & 8 & 8 & 0 & 0 & 0 \\ 4 & 4 & 1 & 0 & 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 2 & 2 & 0 & 0 & 2 & 2 & 0 & 1 \\ 1 & 1 & 17 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

For this problem, you are encouraged to use a computer to compute the eigenvalues/vectors for $\frac{1}{2}(B + B^\top)$.

a) Solve the problem

$$\begin{aligned} &\textit{Maximize:} && x^\top Bx \\ &\textit{Subject to:} && ||x|| = 50 \\ & && x \in \mathbb{R}^{10} \end{aligned}$$

b) Solve the problem

$$\begin{aligned} &\textit{Minimize:} && x^\top Bx \\ &\textit{Subject to:} && ||x|| = 50 \\ & && x \in \mathbb{R}^{10} \end{aligned}$$

c) Solve the problem

$$\begin{aligned} &\textit{Minimize:} && x^\top Bx \\ &\textit{Subject to:} && ||x|| = 25 \\ & && x \in \mathbb{R}^{10} \end{aligned}$$

d) Solve the problem

$$\begin{aligned} &\textit{Maximize:} && |x^\top Bx| \\ &\textit{Subject to:} && ||x|| = 25 \\ & && x \in \mathbb{R}^{10} \end{aligned}$$

EXERCISE 15.19. *(Minimizing $|x^\top Bx|$) Let B be a real $n \times n$ matrix (not necessarily symmetric). Fix r as a positive real number. We want to solve*

$$\text{Minimize:} \quad |x^\top Bx| \tag{15.41}$$

$$\text{Subject to:} \quad ||x|| = r \tag{15.42}$$

$$x \in \mathbb{R}^n \tag{15.43}$$

Define $A = \frac{1}{2}(B + B^\top)$. Define $\lambda_1, \ldots, \lambda_n$ as the (necessarily real) eigenvalues of A, with corresponding orthonormal eigenvectors $\{v_1, \ldots, v_n\}$.

a) Define $A = \frac{1}{2}(B + B^\top)$. Use arguments similar to those in Section 15.6 to show our problem is equivalent to

$$\begin{aligned} &\text{Minimize:} && |x^\top Ax| \\ &\text{Subject to:} && ||x|| = r \\ & && x \in \mathbb{R}^n \end{aligned}$$

b) Consider the case when all eigenvalues of A are nonnegative. Use (15.23) *to solve the problem of (a).*

c) Consider the case when all eigenvalues of A are nonpositive. Use (15.23) *to solve the problem of (a).*

d) Consider the case when at least one eigenvalue of A is positive, and at least one eigenvalue is negative. Use (15.23) *to solve the problem of (a) and show the optimal objective value is zero!*

e) Solve problem (15.41)-(15.43) *to find an optimal x^* in the case $r = 1$ and*

$$B = \begin{bmatrix} 1 & 2 & 4 \\ 1 & 8 & 10 \\ 3 & 3 & 30 \end{bmatrix}$$

You are encouraged to use a computer, such as Matlab, to compute eigenvalues/vectors for $\frac{1}{2}(B + B^\top)$.

f) Solve problem (15.41)-(15.43) *to find an optimal x^* in the case $r = 1$ and*

$$B = \begin{bmatrix} 1 & 3 & -3 & 4 \\ 0 & 2 & 3 & 9 \\ 1 & 0 & -5 & 0 \\ 3 & 0 & 0 & 1 \end{bmatrix}$$

You are encouraged to use a computer to compute eigenvalues/vectors for $\frac{1}{2}(B + B^\top)$.

CHAPTER 16.

Singular values

Fix positive integers m, n. This chapter first develops the singular value decomposition (SVD) of real $m \times n$ matrices. For all but the final section, the field is $\mathbb{R}$, the inner product is $\langle x, y \rangle = x^\top y$ for all $x, y \in \mathbb{R}^n$, the norm is $||x|| = \sqrt{x^\top x}$ for all $x \in \mathbb{R}^n$. The final section uses the same techniques to develop the corresponding SVD results for $m \times n$ matrices with complex values (readers seeking a quick derivation can skip to that section without loss of continuity). Examples in this chapter often display numerical values by using a finite-decimal approximation of an infinite-precision real number that exists and that would make equality hold.

16.1. Rank 1 matrices

Consider the following real matrices D and E:

$$D = \begin{bmatrix} 3 \\ -1 \\ 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 & 2 \end{bmatrix} = \begin{bmatrix} 3 & 3 & 0 & 6 \\ -1 & -1 & 0 & -2 \\ 2 & 2 & 0 & 4 \end{bmatrix}$$

$$E = \begin{bmatrix} 0 \\ 5 \\ 1 \\ 1 \end{bmatrix} \begin{bmatrix} 2 & 3 & 4 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 10 & 15 & 20 \\ 2 & 3 & 4 \\ 2 & 3 & 4 \end{bmatrix}$$

Both D and E are defined as the product of a nonzero column vector and a nonzero row vector. Observe that matrix D has at least one nonzero column, and all columns of D are multiples of its first nonzero column. The same holds for matrix E. *Both D and E are rank 1 matrices*: The dimension of their column space is one.

This holds in general: Fix vectors $c \in \mathbb{R}^m$ and $r \in \mathbb{R}^n$. Assume $c \neq 0$ and $r \neq 0$. Consider the $m \times n$ matrix

$$A = cr^\top \tag{16.1}$$

This is a multiplication of the nonzero column vector c and the nonzero row vector $r^\top$:

$$A = \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_m \end{bmatrix} \begin{bmatrix} r_1 & r_2 & \cdots & r_n \end{bmatrix} = \begin{bmatrix} c_1r_1 & c_1r_2 & \cdots & c_1r_n \\ c_2r_1 & c_2r_2 & \cdots & c_2r_n \\ c_3r_1 & c_3r_2 & \cdots & c_3r_n \\ \vdots & \vdots & \ddots & \vdots \\ c_mr_1 & c_mr_2 & \cdots & c_mr_n \end{bmatrix}$$

We claim that A is a rank 1 matrix. Indeed, observe that for all $x \in \mathbb{R}^n$ we have

$$Ax = c \underbrace{r^\top x}_{\text{scalar}} \in Span\{c\}$$

and so $Col(A) \subseteq Span\{c\}$, which means $\dim Col(A) \leq 1$. On the other hand, A is not the all-zero matrix (see Exercise 16.28 to show $Ar \neq 0$), which means $\dim Col(A) \geq 1$. Therefore, $\dim Col(A) = 1$ and so $\text{rank}(A) = 1$. This proves one direction of the following lemma.

LEMMA 16.1. *A is a real-valued rank 1 matrix of size $m \times n$ if and only if $A = cr^\top$ for some nonzero vectors $c \in \mathbb{R}^m$ and $r \in \mathbb{R}^n$.*

PROOF. The reverse direction is proven in the paragraph preceding the lemma. Exercise 16.2 proves the forward direction. □

It is useful to put the rank 1 matrix $A = cr^\top$ into a standard form by normalizing c and r. Define vectors u and v and a real number σ by

$$u = \frac{c}{||c||}$$
$$v = \frac{r}{||r||}$$
$$\sigma = ||c|| \cdot ||r||$$

The positive scalar σ is called the *singular value* for this rank 1 matrix A. With these

definitions we have $cr^\top = \sigma uv^\top$ and so

$$\boxed{A = \sigma uv^\top} \tag{16.2}$$

The form (16.2), where $\sigma > 0$, $\|u\| = \|v\| = 1$, is called the *singular value decomposed form* of the rank 1 matrix A. From (16.2) we obtain for each $x \in \mathbb{R}^n$

$$\|Ax\| = \left\|\sigma u(v^\top x)\right\| \overset{(a)}{=} |\sigma(v^\top x)| \cdot \underbrace{\|u\|}_{1} = |\sigma| \cdot |v^\top x| \overset{(b)}{\leq} |\sigma| \cdot \underbrace{\|v\|}_{1} \cdot \|x\|$$

where step (a) pulls out a scalar from the norm; step (b) holds by the Cauchy-Schwarz inequality. Therefore

$$\|Ax\| \leq \sigma \|x\| \qquad \forall x \in \mathbb{R}^n \tag{16.3}$$

Thus, for the linear map $T(x) = Ax$, the scalar σ bounds the expansion factor between the norm of the input x and the norm of the output Ax. This is the best possible bound because inequality (16.3) holds with equality whenever x is a scalar multiple of v (notice from (16.2) that $Av = \sigma u$ so $\|Av\| = \sigma$). If we want to maximize the norm of Ax over all input vectors x that have a given norm, then x should point in the direction of v.

16.2. Singular value decomposition (SVD)

In Section 16.7 we show that every real nonzero $m \times n$ matrix A has the following *singular value decomposition* (SVD):

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top \tag{16.4}$$

where $r = \operatorname{rank}(A) \in \{1, \ldots, \min\{m, n\}\}$ and

- $\sigma_1, \ldots, \sigma_r$ are positive real numbers called the *singular values* of A.
- $\{u_1, \ldots, u_r\}$ are orthonormal vectors in $\mathbb{R}^m$.
- $\{v_1, \ldots, v_r\}$ are orthonormal vectors in $\mathbb{R}^n$.

The decomposition (16.4) shows that the rank-r matrix A is a sum of r rank 1 matrices. Every real valued matrix of rank $r \geq 1$ can be written in the form (16.4). (The case $\operatorname{rank}(A) = 0$ occurs if and only if A is the all-zero matrix.)

The decomposition (16.4) can equally be written

$$A = U\Sigma V^\top \tag{16.5}$$

where Σ is a $r \times r$ diagonal matrix with diagonal entries $\sigma_1, \ldots, \sigma_r$; U is a $m \times r$ matrix with columns given by the orthonormal vectors $\{u_1, \ldots, u_r\}$; V is a $n \times r$ matrix with columns given by the orthonormal vectors $\{v_1, \ldots, v_r\}$:

$$\text{(16.6)} \quad \Sigma = \begin{bmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \sigma_r \end{bmatrix} \quad , \quad U = \begin{bmatrix} u_1 & u_2 & \cdots & u_r \end{bmatrix} \quad , \quad V = \begin{bmatrix} v_1 & v_2 & \cdots & v_r \end{bmatrix}$$

Orthonormality of the columns of U and V implies

$$U^\top U = I \sim r \times r \quad , \quad V^\top V = I \sim r \times r$$

which are instances of possibly nonsquare matrices multiplying to the square identity matrix. Equivalence of (16.4) and (16.5) can be seen by expanding the orthonormal vectors $\{v_1, \ldots, v_r\}$ in $\mathbb{R}^n$ to an orthonormal basis $\{v_1, \ldots, v_r, v_{r+1}, \ldots, v_n\}$ for $\mathbb{R}^n$, then observing that defining A by either (16.4) or (16.5) yields the same input-output response on the basis $\{v_1, \ldots, v_n\}$:

$$Av_i = \begin{cases} \sigma_i u_i & \text{if } i \in \{1, \ldots, r\} \\ 0 & \text{if } i \in \{r+1, \ldots, n\} \end{cases}$$

A proof of the decomposition (16.4) is given in Section 16.7.

For example, we present a nonzero 4×3 matrix A in the two different SVD forms (16.4) and (16.5):

$$A = \begin{bmatrix} 5 & 1 & 0 \\ 5 & 1 & 0 \\ 1 & 5 & 0 \\ 1 & 5 & 0 \end{bmatrix} = \underbrace{6\sqrt{2}}_{\sigma_1} \underbrace{\begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}}_{u_1} \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \end{bmatrix}}_{v_1^\top} + \underbrace{4\sqrt{2}}_{\sigma_2} \underbrace{\begin{bmatrix} 1/2 \\ 1/2 \\ -1/2 \\ -1/2 \end{bmatrix}}_{u_2} \underbrace{\begin{bmatrix} 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix}}_{v_2^\top}$$

$$\text{(16.7)} \qquad = \underbrace{\begin{bmatrix} 1/2 & 1/2 \\ 1/2 & 1/2 \\ 1/2 & -1/2 \\ 1/2 & -1/2 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 6\sqrt{2} & 0 \\ 0 & 4\sqrt{2} \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \end{bmatrix}}_{V^\top}$$

so $rank(A) = 2$, the two singular values of A are $\sigma_1 = 6\sqrt{2}$ and $\sigma_2 = 4\sqrt{2}$, and

$$U^\top U = V^\top V = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

This example was created by arbitrarily creating two orthonormal vectors $\{u_1, u_2\}$ in $\mathbb{R}^4$ and two orthonormal vectors $\{v_1, v_2\}$ in $\mathbb{R}^3$ and using positive values of σ_1, σ_2 to ensure the result of $\sum_{i=1}^{2} \sigma_i u_i v_i^\top$ was a nice-looking integer matrix A. If the nice-looking integer matrix A was given *first*, it would not be obvious how to *obtain* the SVD. A general method for obtaining the SVD for a given matrix A is described in a later section.

16.2.1. Extending to $\sigma_1, \ldots, \sigma_n$

Singular values for a real $m \times n$ matrix A are usually ordered from greatest to least, so with $r = \text{rank}(A)$ we have

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

Singular values are always *positive* for indices $i \in \{1, \ldots, r\}$. If $r < n$, it is convenient to define $\sigma_i = 0$ for $i \in \{r+1, \ldots, n\}$. So

$$\begin{aligned} \sigma_i &> 0 \quad \text{if } i \in \{1, \ldots, r\} \\ \sigma_i &= 0 \quad \text{if } i \in \{r+1, \ldots, n\} \end{aligned}$$

In Section 16.8 it is shown that

$$\boxed{\sigma_n \, ||x|| \leq ||Ax|| \leq \sigma_1 \, ||x|| \quad \forall x \in \mathbb{R}^n} \tag{16.8}$$

where σ_1 is the largest singular value and σ_n is the smallest ($\sigma_n = 0$ if and only if $r < n$). Thus, when the $m \times n$ matrix A is viewed as a linear map from $\mathbb{R}^n$ to $\mathbb{R}^m$, the largest singular value σ_1 has the interpretation of being the maximum possible expansion factor between the norm of the input x and the output Ax for the linear map A. The smallest singular value σ_n is the minimum possible expansion factor, and is zero if and only if $r < n$.

The example A matrix of (16.7) has $r = 2$ and $n = 3$ with $\sigma_1 = 6\sqrt{2}$, $\sigma_2 = 4\sqrt{2}$, so we define $\sigma_3 = 0$ and (16.8) yields

$$0 \leq ||Ax|| \leq 6\sqrt{2} \, ||x|| \quad \forall x \in \mathbb{R}^3$$

16.2.2. A stylistic choice for $U\Sigma V^\top$

The $m \times n$ matrix A has rank 0 if and only if it is the all-zero matrix. If A is not the zero matrix, its rank r satisfies

$$1 \le r \le \min\{m, n\}$$

Some people like to do this:

- Extend $\{v_1, ..., v_r\}$ to an orthonormal basis for $\mathbb{R}^n$: $\{v_1, ..., v_n\}$
- Extend $\{u_1, ..., u_r\}$ to an orthonormal basis for $\mathbb{R}^m$: $\{u_1, ..., u_m\}$

Then we can write

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top = \sum_{i=1}^{\min\{m,n\}} \sigma_i u_i v_i^\top$$

The added terms are 0 since they are multiplied by $\sigma_i = 0$ when $i > r$.

Similarly, we can augment the SVD form $A = U\Sigma V^\top$, where Σ is a $r \times r$ diagonal matrix with *positive* diagonal entries, to make Σ a $\min\{m, n\} \times \min\{m, n\}$ diagonal matrix with *nonnegative* diagonal entries, where the added entries are just zero. For example, for the 4×3 matrix A in (16.7) we have $\min\{4, 3\} = 3$ and

$$A = \begin{bmatrix} 5 & 1 & 0 \\ 5 & 1 & 0 \\ 1 & 5 & 0 \\ 1 & 5 & 0 \end{bmatrix} = \underbrace{\begin{bmatrix} 1/2 & 1/2 & ? \\ 1/2 & 1/2 & ? \\ 1/2 & -1/2 & ? \\ 1/2 & -1/2 & ? \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 6\sqrt{2} & 0 & 0 \\ 0 & 4\sqrt{2} & 0 \\ 0 & 0 & 0 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 1/\sqrt{2} & 1/\sqrt{2} & 0 \\ 1/\sqrt{2} & -1/\sqrt{2} & 0 \\ ? & ? & ? \end{bmatrix}}_{V^\top} \tag{16.9}$$

The old Σ matrix in (16.7) was a diagonal 2×2, the new Σ is a diagonal 3×3 but the third singular value is $\sigma_3 = 0$. This requires adding a third column to the old matrix U and a third row to the old matrix $V^\top$. The missing entries, shown as question marks in the equation above, can be filled with *any* real numbers without changing the matrix A. This is because they will be multiplied by 0 in the last row or last column of Σ. It is standard to make u_3 the missing third column of U, and $v_3^\top$ the missing third row of V. Readers are invited to do this in Exercise 16.5. This convention for treating the indices $i \ge r + 1$ should not distract from the fundamental structure (16.4): *Every rank-r matrix has r positive singular values, and can be written as a sum of r rank 1 matrices.*

16.3. Squared singular values as eigenvalues of $A^\top A$ and $AA^\top$

Let A be a real and nonzero $m \times n$ matrix. Define $r = \text{rank}(A)$. Since A is nonzero we know $1 \le r \le \min\{m, n\}$. Lemma 8.5 implies

$$r = \text{rank}(A) = \text{rank}(A^\top) = \text{rank}(A^\top A) = \text{rank}(AA^\top)$$

A later section proves that every real or complex matrix has a SVD. So, assuming A has a SVD, how can we *find* it? Also, is the SVD unique, or are there many SVDs? Start by writing A in the general SVD form

$$A = U\Sigma V^\top \tag{16.10}$$

where Σ is a $r \times r$ diagonal matrix with positive diagonal entries $\sigma_1, \ldots, \sigma_r$, U has orthonormal columns $\{u_1, \ldots, u_r\}$ in $\mathbb{R}^m$, V has orthonormal columns $\{v_1, \ldots, v_r\}$ in $\mathbb{R}^n$. We do not yet know the matrices U, Σ, V. However, orthonormality implies

$$U^\top U = V^\top V = I \quad \text{[the } r \times r \text{ identity matrix]}$$

and so

$$V^\top v_i = e_i \quad \forall i \in \{1, \ldots, r\} \tag{16.11}$$

$$U^\top u_i = e_i \quad \forall i \in \{1, \ldots, r\} \tag{16.12}$$

where $e_i \in \mathbb{R}^r$ is the ith column of I. Observe that $A^\top A$ and $AA^\top$ are both real and symmetric matrices, and (16.10) gives

$$\begin{aligned} A^\top A &= V\Sigma \underbrace{U^\top U}_{I} \Sigma V^\top = V\Sigma^2 V^\top \quad \texttt{[size } n \times n\texttt{]} \\ AA^\top &= U\Sigma \underbrace{V^\top V}_{I} \Sigma U^\top = U\Sigma^2 U^\top \quad \texttt{[size } m \times m\texttt{]} \end{aligned} \tag{16.13}$$

This is similar to a diagonalizable form, although the V and U matrices are not necessarily square. Nevertheless, this form means $\sigma_1^2, \sigma_2^2, \ldots, \sigma_r^2$ are eigenvalues of $A^\top A$ with corresponding eigenvectors $\{v_1, \ldots, v_r\}$. To see this, for each $i \in \{1, \ldots, r\}$ we have

$$\begin{aligned} A^\top A v_i &\overset{(a)}{=} V\Sigma^2 V^\top v_i \\ &\overset{(b)}{=} V\Sigma^2 e_i \\ &\overset{(c)}{=} \sigma_i^2 V e_i \end{aligned}$$

$$\overset{(d)}{=} \sigma_i^2 v_i$$

where (a) holds by substituting (16.13); (b) holds by (16.11); (c) holds because multiplying a matrix on the right by e_i extracts the ith column of that matrix (and the ith column of Σ^2 is $\sigma_i^2 e_i$); (d) holds for the same reason. Thus, σ_i^2 is indeed an eigenvalue of $A^\top A$ with eigenvector v_i (recall v_i is nonzero because it is assumed to have norm 1). A similar argument proves σ_i^2 is also an eigenvalue of $AA^\top$ with eigenvector u_i. Since $A^\top A$ and $AA^\top$ both have rank r, the only other possible eigenvalues of these matrices are zero:

- $A^\top A$ has $n - r$ repeated eigenvalues of 0 (so $\dim Null(A^\top A) = n - r$).
- $AA^\top$ has $m - r$ repeated eigenvalues of 0 (so $\dim Null(AA^\top) = m - r$).

In summary, the r positive values $\sigma_1, \ldots, \sigma_r$ on the diagonal of Σ are precisely the square roots of the positive eigenvalues of $A^\top A$ (and they are also precisely the square roots of the positive eigenvalues of $AA^\top$).[1] *The values $\sigma_1, \ldots, \sigma_r$ are called the singular values of A, and they are found by computing the positive eigenvalues of either $A^\top A$ or $AA^\top$.*

For the SVD $A = U\Sigma V^\top$, the ordering of the singular values $\sigma_1, \ldots, \sigma_r$ along the diagonal entries of Σ is arbitrary. However, it is conventional to list $\sigma_1, \ldots, \sigma_r$ in order from greatest to least (so σ_1 is largest). Under this ordering the $r \times r$ matrix Σ is unique. However, U and V are not unique. For example, the $n \times n$ identity matrix I has many SVDs, two of them are:

$$I = III \quad \texttt{[so } U = V = \Sigma = I\texttt{]}$$
$$I = CIC^\top \quad \texttt{[for any } n \times n \texttt{ matrix } C \texttt{ with orthonormal columns]}$$

Of course, the singular values of I are always $\sigma_1 = \sigma_2 = \cdots = \sigma_n = 1$.

- Example 2×2 matrix A with $\text{rank}(A) = 2$:

$$A = \begin{bmatrix} 3 & 4 \\ -10 & 0 \end{bmatrix} \implies A^\top A = \begin{bmatrix} 109 & 12 \\ 12 & 16 \end{bmatrix}$$

$$\det(A^\top A - \lambda I) = (109 - \lambda)(16 - \lambda) - 12^2 = 0$$

$$\lambda_1 = \sigma_1^2 = 110.5234$$
$$\lambda_2 = \sigma_2^2 = 14.4766$$

$$\sqrt{14.4766}\,||x|| \le ||Ax|| \le \sqrt{110.5234}\,||x|| \quad \forall x \in \mathbb{R}^2$$

[1] A side consequence of this analysis is that all eigenvalues of $A^\top A$ and $AA^\top$ are real and nonnegative, but we already knew that from the fact that both $A^\top A$ and $AA^\top$ are real, symmetric, and positive semidefinite (Lemma 15.6).

- Example 4×2 matrix B with $\text{rank}(B) = 2$:

$$B = \begin{bmatrix} 1 & 3 \\ -7 & -2 \\ 1 & 1 \\ 0 & 8 \end{bmatrix} \implies B^\top B = \begin{bmatrix} 51 & 18 \\ 18 & 78 \end{bmatrix}$$

$$\det(B^\top B - \lambda I) = (51 - \lambda)(78 - \lambda) - 18^2 = 0$$

$$\lambda_1 = \sigma_1^2 = 87$$
$$\lambda_2 = \sigma_2^2 = 42$$

$$\sqrt{42}\,||x|| \le ||Bx|| \le \sqrt{87}\,||x|| \quad \forall x \in \mathbb{R}^2$$

- Example 4×3 matrix C with $\text{rank}(C) = 2$, where column 3 is the sum of the first two columns:

$$C = \begin{bmatrix} 2 & 3 & 5 \\ 2 & 0 & 2 \\ 1 & 3 & 4 \\ 1 & 0 & 1 \end{bmatrix} \implies C^\top C = \begin{bmatrix} 10 & 9 & 19 \\ 9 & 18 & 27 \\ 19 & 27 & 46 \end{bmatrix}$$

$$\det(C^\top C - \lambda I) = -\lambda^3 + 74\lambda^2 - 297\lambda = 0$$

$$\lambda_1 = \sigma_1^2 = 69.7414$$
$$\lambda_2 = \sigma_2^2 = 4.2586$$
$$\lambda_3 = \sigma_3^2 = 0$$

$$0 \le ||Cx|| \le \sqrt{69.7414}\,||x|| \quad \forall x \in \mathbb{R}^3$$

- Example 2×5 matrix D with $\text{rank}(D) = 2$:

$$D = \begin{bmatrix} 6 & -9 & 0 & 0 & 4 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix} \implies DD^\top = \begin{bmatrix} 133 & 1 \\ 1 & 5 \end{bmatrix}$$

where it is more convenient to compute eigenvalues of the 2×2 matrix $DD^\top$ than the 5×5 matrix $D^\top D$:

$$\det(DD^\top - \lambda I) = (133 - \lambda)(5 - \lambda) - 1 = 0$$

$$\lambda_1 = \sigma_1^2 = 133.0078$$
$$\lambda_2 = \sigma_2^2 = 4.9922$$
$$\lambda_3 = \lambda_4 = \lambda_5 = 0$$

$$0 \le ||Dx|| \le \sqrt{133.0078}\,||x|| \quad \forall x \in \mathbb{R}^5$$

16.4. SVD for a rank 1 matrix

In some cases a SVD can be found easily. For example, if we are given a 4×2 rank 1 matrix:

$$A = \begin{bmatrix} 3 \\ -9 \\ 0 \\ 1 \end{bmatrix} \begin{bmatrix} 5 & 2 \end{bmatrix} = \begin{bmatrix} 15 & 6 \\ -45 & -18 \\ 0 & 0 \\ 5 & 2 \end{bmatrix}$$

then the only thing to do is define u_1 and v_1 by normalizing the column and row vectors:

$$u_1 = \frac{1}{\sqrt{91}} \begin{bmatrix} 3 \\ -9 \\ 0 \\ 1 \end{bmatrix}, v_1^\top = \frac{1}{\sqrt{29}} \begin{bmatrix} 5 & 2 \end{bmatrix} \Longrightarrow \boxed{A = \sqrt{2639}\, u_1 v_1^\top} \tag{16.14}$$

where we have used $\sqrt{91 \cdot 29} = \sqrt{2639}$. Observe that $||u|| = ||v|| = 1$. The representation of A in (16.14) has the form $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$ in the special case $r = 1$, so the sum only has one term and there is only one singular value:

$$\boxed{\sigma_1 = \sqrt{2639}}$$

This also has the form $U\Sigma V^\top$ where Σ is the 1×1 matrix $(\sqrt{2639})$, so $A = u_1\,(\sqrt{2639})\,v_1^\top$. Since $r = 1$ and $n = 2$, we define $\sigma_2 = 0$ and obtain

$$0 \le ||Ax|| \le \sqrt{2639}\,||x|| \quad \forall x \in \mathbb{R}^2$$

Readers who observe that $\min\{4, 2\} = 2$ may prefer the SVD structure:

$$A = \underbrace{\begin{bmatrix} u_1 & ? \end{bmatrix}}_{U_{new}} \underbrace{\begin{bmatrix} \sqrt{2639} & 0 \\ 0 & 0 \end{bmatrix}}_{\Sigma_{new}} \underbrace{\begin{bmatrix} v_1^\top \\ ? \end{bmatrix}}_{V_{new}^\top}$$

where u_1, v_1 are defined in (16.14); the second column of U_{new} is any column vector $u_2 \in \mathbb{R}^4$ such that $\|u_2\| = 1$ and $u_2^\top u_1 = 0$; the second row of $V_{new}^\top$ is any row vector $v_2^\top \in \mathbb{R}^2$ such that $\|v_2\| = 1$ and $v_2^\top v_1 = 0$. This convention requires finding such u_2, v_2 vectors, but of course these vectors have no impact because they are multiplied by 0 in the above equation for A.

16.5. General procedure for computing SVD

Given a real and nonzero $m \times n$ matrix A, we can find a SVD $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ by carrying out the following procedure:

1. Find the nonzero eigenvalues $\lambda_1, \ldots, \lambda_r$ of $A^\top A$, where $r = \text{rank}(A)$. These are possibly repeated according to algebraic multiplicity. Define $\sigma_i = \sqrt{\lambda_i}$ for $i \in \{1, \ldots, r\}$.
   A correct answer automatically has *positive* eigenvalues $\lambda_i > 0$ for $i \in \{1, \ldots, r\}$.
2. Find corresponding orthogonal eigenvectors of $A^\top A$ and normalize them to define $\{v_1, \ldots, v_r\}$.
   When v_i, v_j have a repeated eigenvalue (so $\lambda_i = \lambda_j$), care must be taken to ensure $v_i^\top v_j = 0$.
3. Define $u_i = \frac{Av_i}{\sigma_i}$ for $i \in \{1, \ldots, r\}$.
   A correct answer automatically has $\|u_i\| = 1$ and $u_i^\top u_j = 0$ for all $i \neq j$.

Note that $A^\top A$ is real and symmetric. Step 2 in the above procedure is easiest when all nonzero eigenvalues $\lambda_1, \ldots, \lambda_r$ of $A^\top A$ are distinct, in which case *any* eigenvectors v_i found for each λ_i will *automatically* have the orthogonality property $v_i^\top v_j = 0$ for all $i \neq j$. If there is an eigenvalue λ_i that has an algebraic multiplicity $m_i > 1$, then $\dim Null(A^\top A - \lambda_i I) = m_i$ (since the real matrix is symmetric) and we must find (any) m_i orthonormal vectors in $Null(A^\top A - \lambda_i I)$. These will *automatically* be orthogonal to eigenvectors associated with eigenvalues $\lambda_j \neq \lambda_i$ because $Null(A^\top A - \lambda_i I)$ and $Null(A^\top A - \lambda_j I)$ can be shown to be orthogonal subspaces (see Exercise 15.7). The total number of nonzero eigenvalues $\sigma_1, \ldots, \sigma_r$ (counted with algebraic multiplicity) is the rank of A. One can also find the rank by using $r = \text{rank}(A) = \text{rank}(A^\top A) = \text{rank}(AA^\top) = \text{rank}(A^\top)$. The singular values are typically placed in order:

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

which of course means the vectors $v_i \in \mathbb{R}^n$ and $u_i \in \mathbb{R}^m$ should be indexed according to this order. The form $A = U\Sigma V^\top$ is achieved by using (16.6). Remember that V is defined in (16.6), but its *transpose* is used in the formula $A = U\Sigma V^\top$.

16.6. Example

Consider the 3×2 matrix

$$A = \begin{bmatrix} 7 & -3 \\ 2 & 1 \\ 0 & 3 \end{bmatrix}$$

This is a rank 2 matrix. We want to find an SVD for A:

$$A = \sigma_1 u_1 v_1^\top + \sigma_2 u_2 v_2^\top = U \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix} V^\top$$

We have

$$A^\top A = \begin{bmatrix} 53 & -19 \\ -19 & 19 \end{bmatrix}$$

The eigenvalues of $A^\top A$ solve $\det(A^\top A - \lambda I) = 0$ and are

$$\lambda_1 = 36 + 5\sqrt{26} \implies \sigma_1 = \sqrt{\lambda_1} = 7.8419$$
$$\lambda_2 = 36 - 5\sqrt{26} \implies \sigma_2 = \sqrt{\lambda_2} = 3.2411$$

The larger eigenvalue has been intentionally labeled λ_1 to ensure $\sigma_1 \geq \sigma_2$. From this information alone we can assert (using (16.8) and the fact $r = n = 2$ here):

$$\boxed{3.2411\,||x|| \leq ||Ax|| \leq 7.8419\,||x|| \qquad \forall x \in \mathbb{R}^2}$$

The eigenvectors are nonzero vectors in $Null(A^\top A - \lambda I)$ for $\lambda \in \{\lambda_1, \lambda_2\}$:

$$A - \lambda I = \begin{bmatrix} 53 - \lambda & -19 \\ -19 & 19 - \lambda \end{bmatrix} \to \begin{bmatrix} -19 & 19 - \lambda \\ 53 - \lambda & -19 \end{bmatrix} \to \begin{bmatrix} -19 & 19 - \lambda \\ 0 & 0 \end{bmatrix}$$

an eigenvector $z = (a, 1)$ must satisfy

$$-19a + 19 - \lambda = 0 \implies a = 1 - \frac{\lambda}{19}$$

so

$$\lambda_1 = 36 + 5\sqrt{26} \implies z_1 = \begin{bmatrix} 1 - \lambda_1/19 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{-17-5\sqrt{26}}{19} \\ 1 \end{bmatrix} \implies v_1 = \begin{bmatrix} -0.9129 \\ 0.4082 \end{bmatrix}$$

$$\lambda_2 = 36 - 5\sqrt{26} \implies z_2 = \begin{bmatrix} 1 - \lambda_2/19 \\ 1 \end{bmatrix} = \begin{bmatrix} \frac{-17+5\sqrt{26}}{19} \\ 1 \end{bmatrix} \implies v_2 = \begin{bmatrix} 0.4082 \\ 0.9129 \end{bmatrix}$$

One can verify that z_1 and z_2 are indeed orthogonal, and v_1, v_2 are normalized versions of z_1, z_2. Then

$$q_1 = Av_1 = \begin{bmatrix} -7.6149 \\ -1.4176 \\ 1.2245 \end{bmatrix} \implies u_1 = \begin{bmatrix} -0.9710 \\ -0.1808 \\ 0.1561 \end{bmatrix}$$

$$q_2 = Av_2 = \begin{bmatrix} 0.1185 \\ 1.7292 \\ 2.7387 \end{bmatrix} \implies u_2 = \begin{bmatrix} 0.0366 \\ 0.5335 \\ 0.8450 \end{bmatrix}$$

One can verify that $\|q_1\| = \sigma_1$ and $\|q_2\| = \sigma_2$, that q_1, q_2 are orthogonal, and u_1, u_2 are normalized versions of q_1, q_2. Finally

$$A = \sigma_1 \begin{bmatrix} -0.9710 \\ -0.1808 \\ 0.1561 \end{bmatrix} \begin{bmatrix} -0.9129 & 0.4082 \end{bmatrix} + \sigma_2 \begin{bmatrix} 0.0366 \\ 0.5335 \\ 0.8450 \end{bmatrix} \begin{bmatrix} 0.4082 & 0.9129 \end{bmatrix} = \begin{bmatrix} 7.0000 & -3.0000 \\ 2.0000 & 1.0000 \\ 0.0000 & 3.0000 \end{bmatrix}$$

An equivalent representation is

$$A = \underbrace{\begin{bmatrix} -0.9710 & 0.0366 \\ -0.1808 & 0.5335 \\ 0.1561 & 0.8450 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 7.8419 & 0 \\ 0 & 3.2411 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} -0.9129 & 0.4082 \\ 0.4082 & 0.9129 \end{bmatrix}}_{V^\top}$$

16.7. Proof of SVD existence

Fix A as a nonzero $m \times n$ matrix. Let $r = \text{rank}(A)$ and note that $r \in \{1, \ldots, \min\{m, n\}\}$. We show that A has the SVD (16.4). Note that $A^\top A$ is a real, symmetric, positive semidefinite $n \times n$ matrix (Lemma 15.6), so it has an orthonormal basis of eigenvectors with real and nonnegative eigenvalues. Specifically, let $\{v_1, \ldots, v_n\}$ be the orthonormal basis for $\mathbb{R}^n$ that are eigenvectors of $A^\top A$ with corresponding eigenvalues $\lambda_i \geq 0$, so

$$A^\top A v_i = \lambda_i v_i \quad \forall i \in \{1, \ldots, n\} \tag{16.15}$$

Now define q_i by

$$q_i = Av_i \quad \forall i \in \{1, \ldots, n\}$$

For convenience, assume the indices are labeled to place the norms of the $\{q_i\}$ vectors in decreasing order, so

$$\|q_1\| \geq \|q_2\| \geq \cdots \geq \|q_n\|$$

Claim 1: $\{q_1, \ldots, q_n\}$ span $Col(A)$.

To prove Claim 1, fix $y \in Col(A)$. Then $y = Ax$ for some $x \in \mathbb{R}^n$. Since $\{v_1, \ldots, v_n\}$ are a basis for $\mathbb{R}^n$ we know

$$x = \sum_{i=1}^{n} a_i v_i$$

for some real numbers a_i. Multiplying both sides of the above equation by A gives

$$\underbrace{Ax}_{y} = \sum_{i=1}^{n} a_i \underbrace{Av_i}_{q_i}$$

and so every $y \in Col(A)$ is a linear combination of $\{q_1, \ldots, q_n\}$, proving Claim 1.

Claim 2: $A = \sum_{j=1}^{n} q_j v_j^\top$.

To prove Claim 2, fix $i \in \{1, \ldots, n\}$. Then

$$\left(\sum_{j=1}^{n} q_j v_j^\top \right) v_i = \sum_{j=1}^{n} q_j v_j^\top v_i = q_i v_i^\top v_i = q_i \|v_i\|^2 = q_i$$

which uses $v_j^\top v_i = 0$ if $i \neq j$, and $\|v_i\| = 1$. This holds for all $i \in \{1, \ldots, n\}$, so the input-output response of the matrix A on the basis $\{v_1, \ldots, v_n\}$ for $\mathbb{R}^n$ is the same as the input-output response of the matrix $\sum_{j=1}^{n} q_j v_j^\top$. They must be the same matrix, which proves Claim 2.

Claim 3: $\{q_1, \ldots, q_n\}$ are orthogonal vectors in $\mathbb{R}^m$.

To prove Claim 3, fix $i, j \in \{1, \ldots, n\}$ with $i \neq j$. Then

$$q_i^\top q_j = (Av_i)^\top (Av_j) = v_i^\top A^\top A v_j \overset{(a)}{=} v_i^\top (\lambda_j v_j) \overset{(b)}{=} \lambda_j v_i^\top v_j = 0$$

where (a) holds by (16.15); (b) holds because v_i and v_j are known to be orthogonal. This proves Claim 3.

Since $\{q_1, \ldots, q_n\}$ are orthogonal vectors in $Col(A)$, *at most* r of them are nonzero (else we would have $r+1$ linearly independent vectors in the r-dimensional space $Col(A)$, which

is impossible). On the other hand, *at least* r of the vectors in $\{q_1, ..., q_n\}$ are nonzero because these vectors span the r-dimensional space $Col(A)$. It follows that exactly r of these vectors are nonzero, and since they are in order, we have

$$||q_i||^2 > 0 \text{ if } i \in \{1, ..., r\} \tag{16.16}$$

$$||q_i||^2 = 0 \text{ if } i \in \{r+1, ..., n\} \tag{16.17}$$

Now define $u_i \in \mathbb{R}^m$ and $\sigma_i \in \mathbb{R}$ by

$$u_i = \frac{q_i}{||q_i||} \quad , \sigma_i = ||q_i|| \quad \forall i \in \{1, ..., r\} \tag{16.18}$$

Since $\{q_1, ..., q_r\}$ are already known to be orthogonal, we know $\{u_1, ..., u_r\}$ are orthonormal. We already know $\{v_1, ..., v_r\}$ are orthonormal. Finally,

$$A \overset{(a)}{=} \sum_{i=1}^{n} q_i v_i^\top \overset{(b)}{=} \sum_{i=1}^{r} q_i v_i^\top \overset{(c)}{=} \sum_{i=1}^{r} ||q_i|| \frac{q_i}{||q_i||} v_i^\top \overset{(d)}{=} \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

where (a) holds by Claim 2; (b) holds by (16.17); (c) is valid because $||q_i|| \neq 0$ by (16.16); (d) holds by definition of u_i and σ_i in (16.18). This completes the proof of the SVD for A.

16.8. Maximum expansion factor

Let A be a real and nonzero $m \times n$ matrix. Suppose A has rank r. Then $r \in \{1, ..., \min\{m, n\}\}$ and we have the SVD

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$$

Assume the singular values have been ordered from greatest to least, so

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

It is often convenient to extend the singular values all the way to n, so $\sigma_i = 0$ for $i \in \{r+1, ..., n\}$, and $\{v_1, ..., v_r\}$ are extended to an orthonormal basis $\{v_1, ..., v_n\}$ for $\mathbb{R}^n$. Exercise 16.1 shows that if $r < n$, then $\{v_{r+1}, ..., v_n\}$ are a basis for $Null(A)$. Every $x \in \mathbb{R}^n$ can be written

$$x = \sum_{j=1}^{n} a_j v_j$$

for some real numbers $a_1, ..., a_n$. So orthonormality of $\{v_1, ..., v_n\}$ implies (by Lemma 8.6b):

$$||x||^2 = \left|\left|\sum_{j=1}^{n} a_j v_j\right|\right|^2 = \sum_{j=1}^{n} a_j^2$$

and

$$Ax = \left(\sum_{i=1}^{r} \sigma_i u_i v_i^\top\right)\left(\sum_{j=1}^{n} a_j v_j\right) = \sum_{i=1}^{r}\sum_{j=1}^{n} \sigma_i a_j u_i v_i^\top v_j = \sum_{i=1}^{r} \sigma_i a_i u_i$$

which uses $v_i^\top v_j = 0$ if $i \neq j$, so orthonormality of $\{u_1, \ldots, u_r\}$ implies

$$||Ax||^2 = \sum_{i=1}^{r} \sigma_i^2 a_i^2$$

It follows that for all $x \in \mathbb{R}^n$ we have $x = \sum_{i=1}^{n} a_i x_i$ for some real numbers $a_1, \ldots, a_n$, which means $||x||^2 = \sum_{i=1}^{n} a_i^2$ and

$$||Ax||^2 = \sum_{i=1}^{r} \sigma_i^2 a_i^2 \leq \sigma_1^2 \sum_{i=1}^{r} a_i^2 \leq \sigma_1^2 ||x||^2 \tag{16.19}$$

$$\boxed{||Ax|| \leq \sigma_1 ||x|| \quad \forall x \in \mathbb{R}^n} \tag{16.20}$$

The largest singular value σ_1 can be viewed as the largest expansion factor in the norm of the input and output in the linear map A. Specifically, consider the optimization problem that searches for a vector that maximizes the expansion factor:

$$\text{Maximize:} \quad \frac{||Ax||}{||x||} \tag{16.21}$$

$$\text{Subject to:} \quad x \in \mathbb{R}^n, x \neq 0 \tag{16.22}$$

By (16.20), the objective value for this problem can be no larger than σ_1. In fact, σ_1 (the largest singular value of A) is the optimal objective for this problem, and it is achieved by the corresponding vector v_1:

$$Av_1 = \sigma_1 u_1 \implies \frac{||Av_1||}{||v_1||} = \frac{\sigma_1 ||u_1||}{||v_1||} = \sigma_1$$

which uses the fact that $||v_1|| = 1$ and $||u_1|| = 1$. Every nonzero multiple of v_1 also achieves this same maximum expansion ratio. Indeed, for a real number $\theta \neq 0$ we have

$$\frac{||A(\theta v_1)||}{||\theta v_1||} = \frac{||Av_1||}{||v_1||} = \sigma_1$$

A closely related problem is

$$\begin{aligned} &\text{Maximize:} && ||Ax|| \\ &\text{Subject to:} && ||x|| = 1 \end{aligned}$$

This new problem is equivalent to problem (16.21)-(16.22) in the following sense: The two problems have the same optimal objective value σ_1; any solution x^* to this new problem is also a solution to (16.21)-(16.22); any solution x^* to (16.21)-(16.22) can be normalized to create a solution to this new problem. This new problem is often preferred to (16.21)-(16.22) because the constraint set is

$$\{x \in \mathbb{R}^n : ||x|| = 1\}$$

which is a closed and bounded subset of $\mathbb{R}^n$. This set is equal to $\{x \in \mathbb{R}^n : ||x||^2 = 1\}$. Also note that finding an x to maximize $||Ax||$ is equivalent to finding an x to maximize $||Ax||^2$. This is useful because derivative and Lagrange multiplier methods can also be used to maximize $||Ax||^2$ subject to $||x||^2 = 1$ to arrive at the optimal objective value σ_1^2.

The problem of *minimizing* the expansion factor is

$$\begin{aligned} \text{Minimize:} \quad & ||Ax|| \\ \text{Subject to:} \quad & ||x|| = 1 \end{aligned}$$

If $Null(A)$ is nontrivial, then any nonzero vector in $Null(A)$ can be normalized to produce a solution to this minimization problem. With $r = \text{rank}(A)$ we have $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ and we can take the orthonormal vectors $\{v_1, ..., v_r\}$ and extend them to an orthonormal basis $\{v_1, ..., v_n\}$ for $\mathbb{R}^n$ with $||Av_i|| = \sigma_i$ for $i \in \{1, ..., n\}$, where

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$$

and $\sigma_i > 0$ for $i \in \{1, ..., r\}$ and $\sigma_i = 0$ for $i > r$. Then an argument similar to (16.19) shows that v_n solves this problem with optimal objective value σ_n, so

$$\sigma_n ||x|| \leq ||Ax|| \leq \sigma_1 ||x|| \quad \forall x \in \mathbb{R}^n$$

For example, in the 3×2 example of Section 16.6 we obtained the SVD:

$$A = \begin{bmatrix} 7 & -3 \\ 2 & 1 \\ 0 & 3 \end{bmatrix} = \underbrace{\begin{bmatrix} -0.9710 & 0.0366 \\ -0.1808 & 0.5335 \\ 0.1561 & 0.8450 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 7.8419 & 0 \\ 0 & 3.2411 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} -0.9129 & 0.4082 \\ 0.4082 & 0.9129 \end{bmatrix}}_{V^\top}$$

and so

$$3.2411 ||x|| \leq ||Ax|| \leq 7.8419 ||x|| \quad \forall x \in \mathbb{R}^2$$

where the upper bound is achieved by any multiple of $v_1 = (-.9129, 0.4082)$ and the lower bound is achieved by any multiple of $v_2 = (0.4082, 0.9129)$.

16.9. Singular values for symmetric matrices

Suppose A is a real and symmetric $n \times n$ matrix. We already know that A has real eigenvalues $\lambda_1, \ldots, \lambda_n$ and corresponding eigenvectors $\{v_1, \ldots, v_n\}$ that are an orthonormal basis of $\mathbb{R}^n$, and A has the diagonalizable form

$$A = S \Lambda S^\top$$

where columns of S are the orthonormal eigenvectors $\{v_1, \ldots, v_n\}$ and Λ is a diagonal matrix of the real eigenvalues. If the eigenvalues are all nonnegative (which occurs when A is symmetric and positive semidefinite) then $A = S \Lambda S^\top$ would be the singular value decomposition! Therefore, if A is a real and symmetric matrix that is positive semidefinite, its singular values are the same as its eigenvalues:

$$\sigma_i = \lambda_i \quad \forall i \in \{1, \ldots, n\}$$

One must be careful to place the eigenvalues in order from greatest to least to respect the conventional ordering for singular values. Once they are in order, we have $A = U \Sigma V^\top$ with $U = V = S$ and $\Sigma = \Lambda$. The form $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ can be obtained by using only the positive eigenvalues of A (there will be exactly r of them, where $r = \text{rank}(A)$).

What if we obtain the diagonalizable form $A = S \Lambda S^\top$ but some of the eigenvalues on the diagonal of Λ are negative? In that case, it turns out the singular value decomposition is $A = U \Sigma V^\top$ where

$$V = S, \quad \Sigma = |\Lambda|, \quad U = \tilde{S}$$

where $|\Lambda|$ is a diagonal matrix with entries equal to the absolute values of Λ, and $\tilde{S}$ is the matrix formed by starting with S and, for each $i \in \{1, \ldots, n\}$ such that $\lambda_i < 0$, multiplying column i of S by -1. Thus, if A is a real and symmetric matrix, its singular values are the absolute values of its eigenvalues:

$$\sigma_i = |\lambda_i| \quad \forall i \in \{1, \ldots, n\}$$

For example, in Section 15.8 we found the diagonalizable form of the following real and symmetric matrix:

$$A = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5257 & -0.8507 \\ 0 & 0.8507 & 0.5257 \end{bmatrix}}_{S} \underbrace{\begin{bmatrix} -2 & 0 & 0 \\ 0 & 4.6180 & 0 \\ 0 & 0 & 2.3820 \end{bmatrix}}_{\Lambda} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5257 & -0.8507 \\ 0 & 0.8507 & 0.5257 \end{bmatrix}^\top}_{S^\top}$$

This matrix A is not positive semidefinite because it has -2 as an eigenvalue. However, we immediately see its singular values are 2, 4.6180, 2.3820 and its SVD is

$$A = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} = \underbrace{\begin{bmatrix} -1 & 0 & 0 \\ 0 & 0.5257 & -0.8507 \\ 0 & 0.8507 & 0.5257 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 2 & 0 & 0 \\ 0 & 4.6180 & 0 \\ 0 & 0 & 2.3820 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 0.5257 & -0.8507 \\ 0 & 0.8507 & 0.5257 \end{bmatrix}^{\top}}_{V^{\top}}$$

where Σ is obtained from Λ by multiplying the first column by -1; U is obtained from S by multiplying the first column by -1. *Oops, this is a correct equation, but we have not placed the singular values in order!* Okay, for this, we must relabel the indices:

$$\texttt{Index1} \to \texttt{Index3} \to \texttt{Index2} \to \texttt{Index1}$$

So permuting the columns of U and V in this way, and similarly permuting the diagonal entries of Σ, we obtain

$$A = \begin{bmatrix} -2 & 0 & 0 \\ 0 & 3 & 1 \\ 0 & 1 & 4 \end{bmatrix} = \underbrace{\begin{bmatrix} 0 & 0 & -1 \\ 0.5257 & -0.8507 & 0 \\ 0.8507 & 0.5257 & 0 \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} 4.6180 & 0 & 0 \\ 0 & 2.3820 & 0 \\ 0 & 0 & 2 \end{bmatrix}}_{\Sigma} \underbrace{\begin{bmatrix} 0 & 0 & 1 \\ 0.5257 & -0.8507 & 0 \\ 0.8507 & 0.5257 & 0 \end{bmatrix}^{\top}}_{V^{\top}}$$

To avoid this ordering mistake, it may be better to work directly with the eigenvalues $\lambda_1, \ldots, \lambda_n$ and eigenvectors $\{v_1, \ldots, v_n\}$ of the real and symmetric matrix A, first put *those* in order of largest-to-smallest $|\lambda_i|$, then create the decompositions $A = S\Lambda S^{\top} = U\Sigma V^{\top}$ using that desirable order.

16.10. Matrix approximation

Suppose A is a real $m \times n$ matrix with a large rank $r \in \{1, \ldots, \min\{m, n\}\}$. Then A has the SVD

$$A = \sum_{i=1}^{r} \sigma_i u_i v_i^{\top} \tag{16.23}$$

Assume the singular values are ordered so that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$$

The SVD (16.23) is an exact representation of the matrix A. To deliver the exact matrix A to someone, one could either transmit the nm real-valued coefficients of A over a communication channel, or transmit $\sigma_1, \ldots, \sigma_r, u_1, \ldots, u_r, v_1, \ldots, v_r$. In many cases we have $r = \min\{m, n\}$, and it is easier to just send the nm coefficients.

Now fix $k \in \{1, \ldots, r-1\}$ and consider the *approximation*

$$\tilde{A} = \sum_{i=1}^{k} \sigma_i u_i v_i^{\top}$$

This approximation includes only information related to the k most significant singular values. If k is much less than r, then $\tilde{A}$ is much easier to represent and communicate: We need only give someone the k most significant singular values $\sigma_1, \ldots, \sigma_k$ and the corresponding $u_1, \ldots, u_k$ and $v_1, \ldots, v_k$ vectors. Often $A \approx \tilde{A}$ even when k is much smaller than r, so we can send much less information than mn coefficients. The approximation gets better and better as we add information related to the additional singular values.

16.11. Singular values for complex matrices

The development of singular values for matrices with complex entries is similar. Let A be a nonzero $m \times n$ matrix with complex entries. Define $r = \text{rank}(A)$, so $1 \le r \le \min\{m, n\}$. We claim A has the following singular value decomposition (SVD)

$$A = U\Sigma V^H = \sum_{i=1}^{r} \sigma_i u_i v_i^H$$

where Σ is a $r \times r$ diagonal matrix with real and positive entries $\sigma_1, \ldots, \sigma_r$ (the positive singular values of A), U is a complex-valued $m \times r$ matrix with orthonormal columns $\{u_1, \ldots, u_r\}$, V is a complex-valued $n \times r$ matrix with orthonormal columns $\{v_1, \ldots, v_r\}$. Orthonormality implies

$$U^H U = I \quad , \quad V^H V = I$$

First assume A has this SVD form. Then

$$A^H A = (U\Sigma V^H)^H (U\Sigma V^H) = V\Sigma \underbrace{U^H U}_{I} \Sigma V^H = V\Sigma^2 V^H$$

$$AA^H = (U\Sigma V^H)(U\Sigma V^H)^H = U\Sigma \underbrace{V^H V}_{I} \Sigma U^H = U\Sigma^2 U^H$$

For each $i \in \{1, \ldots, r\}$, orthonormality of the columns of U and V implies

$$V^H v_i = e_i$$

$$U^H u_i = e_i$$

where e_i is the ith column of the $r \times r$ identity matrix. Using this implies:

- σ_i^2 is an eigenvalue of $A^H A$ with eigenvector v_i.
- σ_i^2 is an eigenvalue of AA^H with eigenvector u_i.
- $Av_i = \sigma_i u_i$

which means that computing singular values $\sigma_1, ..., \sigma_r$ for complex-valued matrices is similar to computing them for real matrices: Just compute the nonzero eigenvalues of $A^H A$ (or AA^H). Since both $A^H A$ and AA^H are Hermitian, their eigenvalues are always real. Further, it can be shown that their eigenvalues are always greater than or equal to zero. The positive ones are the singular values of A. To show that every eigenvalue of $A^H A$ is nonnegative, let (λ, v) be an eigenpair of $A^H A$. Then

$$0 \leq ||Av||^2 = (Av)^H (Av) = v^H (A^H Av) = v^H (\lambda v) = \lambda ||v||^2$$

and so $\lambda \geq 0$. A similar proof shows that every eigenvalue of AA^H is nonnegative.

We now show that every nonzero $m \times n$ matrix A has this singular value decomposition. For simplicity, our goal is to show the form $A = \sum_{i=1}^{r} \sigma_i u_i v_i^H$. First note that $A^H A$ is $n \times n$ Hermitian and hence has real eigenvalues $\lambda_1, ..., \lambda_n$ with corresponding eigenvectors $\{v_1, ..., v_n\}$ that are an orthonormal basis for $\mathbb{C}^n$. For each i define

$$q_i = Av_i$$
$$\sigma_i = ||q_i||$$

For convenience, assume the indices are labeled so that

$$||q_1|| \geq ||q_2|| \geq ... \geq ||q_n|| \geq 0$$

Claim 1: $\{q_1, ..., q_n\}$ are orthogonal.
To prove Claim 1, for $i \neq j$ we have

$$q_i^H q_j = (Av_i)^H (Av_j) = v_i^H (A^H Av_j) = v_i^H (\lambda_j v_j) = \lambda_j (v_i^H v_j) = 0$$

where the last equality holds because v_i and v_j are orthonormal. This proves Claim 1.

Claim 2: $A = \sum_{i=1}^{n} q_i v_i^H$.
To prove Claim 2, we show the input-output behavior of the matrices A and $\sum_{i=1}^{n} q_i v_i^H$ match on the basis $\{v_1, ..., v_n\}$ for $\mathbb{C}^n$. Fix $j \in \{1, ..., n\}$. By definition of q_j, we have $Av_j = q_j$.

Also

$$\left(\sum_{i=1}^{n} q_i v_i^H\right) v_j = q_j v_j^H v_j = q_j$$

which uses $v_j^H v_j = \left\|v_j\right\|^2 = 1$. This proves Claim 2.

Claim 3: We have $||q_i|| > 0$ for $i \in \{1, ..., r\}$ and $||q_i|| = 0$ for $i \in \{r+1, ..., n\}$.

To prove Claim 3, since the $\{q_1, ..., q_n\}$ vectors are orthogonal vectors in $Col(A)$, the nonzero ones are linearly independent. There can be no more than r linearly independent vectors in the r-dimensional subspace $Col(A)$, and so *at most* r of the $\{q_i\}$ vectors are nonzero. On the other hand, since $\{v_1, ..., v_n\}$ form a basis for $\mathbb{C}^n$, it is straightforward to show that $Col(A) = Span\{q_1, ..., q_n\}$ (every $y \in Col(A)$ has the form $y = Ax$ for some $x \in \mathbb{C}^n$), so *at least* r of the $\{q_i\}$ vectors are nonzero. Since these vectors are in order of decreasing norm, the first r of them must be nonzero and the remaining ones are zero, proving Claim 3.

Now we define orthonormal vectors $\{u_1, ..., u_r\}$ by

$$u_i = \frac{q_i}{||q_i||} \quad \forall i \in \{1, ..., r\}$$

and it follows that

$$A \overset{(a)}{=} \sum_{i=1}^{n} q_i v_i^H \overset{(b)}{=} \sum_{i=1}^{r} q_i v_i^H = \sum_{i=1}^{r} ||q_i|| \frac{q_i}{||q_i||} v_i^H \overset{(c)}{=} \sum_{i=1}^{r} \sigma_i u_i v_i^H$$

where (a) holds by Claim 2; (b) holds by Claim 3; (c) holds by definition of u_i and σ_i. We have produced the SVD for the complex-valued matrix A.

16.12. Singular values for A and A^k

Given a complex $m \times n$ matrix A, let $||A||$ denote its largest singular value:

$$||A|| = \max_{x \in \mathbb{C}^n : ||x|| = 1} \{||Ax||\}$$

Then

$$||Ax|| \le ||A|| \cdot ||x|| \quad \forall x \in \mathbb{C}^n$$

If A is a square matrix (size $n \times n$) it holds by induction that

$$\left\|A^k x\right\| \le ||A||^k ||x|| \quad \forall k \in \{0, 1, 2, ...\}, \forall x \in \mathbb{C}^n \tag{16.24}$$

and so

$$\left|\left|A^k\right|\right| \leq ||A||^k \quad \forall k \in \{0, 1, 2, ...\}$$

Given a complex $n \times n$ matrix A, we can define

$$\rho(A) = \inf_{k \in \mathbb{N}} \{\left|\left|A^k\right|\right|^{1/k}\}$$

and we have

$$0 \leq \rho(A) \leq ||A||$$

It is not hard to show that if A is a normal matrix then $||A||$ is the magnitude of its largest-magnitude eigenvalue, and $\rho(A) = ||A||$. In general, $\rho(A)$ can be smaller than $||A||$. For example, consider

$$A = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix} \implies A^m = \begin{bmatrix} 1 & m \\ 0 & 1 \end{bmatrix}$$

It can be shown that $||A|| = 1.6180$, $||A^m|| \geq 1$ for all m, and $||A^m||^{1/m} \to 1$, so $\rho(A) = 1$.

16.13. Exercises

EXERCISE 16.1. *(Singular values and null space) Let A be a $m \times n$ matrix with rank r. Assume $1 \leq r < n$. Let $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ be the SVD, where $\{u_1, ..., u_r\}$ are orthonormal in $\mathbb{R}^m$ and $\{v_1, ..., v_r\}$ are orthonormal in $\mathbb{R}^n$. Extend $\{v_1, ..., v_r\}$ to an orthonormal basis for $\mathbb{R}^n$: $\{v_1, ..., v_r, v_{r+1}, ...v_n\}$.*

a) Compute Av_i and $||Av_i||$ for all $i \in \{1, ..., n\}$.

b) Show that $\{v_{r+1}, ..., v_n\}$ is an orthonormal basis for Null(A).

c) Show that $\{u_1, ..., u_r\}$ is an orthonormal basis for Col(A).

d) Find an $x \in \mathbb{R}^n$ that solves

$$\begin{aligned} &\textit{Maximize:} && ||Ax|| \\ &\textit{Subject to:} && ||x|| = 1 \end{aligned}$$

e) Find an $x \in \mathbb{R}^n$ that solves

$$\begin{aligned} &\textit{Minimize:} && ||Ax|| \\ &\textit{Subject to:} && ||x|| = 1 \end{aligned}$$

EXERCISE 16.2. *(Rank 1 matrices) Fix positive integers m, n. Let A be a real-valued $m \times n$ matrix with rank 1. We want to prove that $A = cr^\top$ for some nonzero vectors $c \in \mathbb{R}^m$ and $r \in \mathbb{R}^n$.*

a) What is the dimension of Null(A)?

b) Explain why there exists a nonzero vector $w \in \mathbb{R}^n$ such that $w \in Null(A)^\perp$.

c) Define $c = Aw$. Explain why $c \neq 0$.

d) Define $r = w/||w||^2$. Prove that $cr^\top = A$. Hint: It suffices to prove the input-output behavior of the matrices $cr^\top$ and A match on a basis of $\mathbb{R}^n$.

EXERCISE 16.3. *(Columns of a rank 1 matrix) Define*

$$A = \begin{bmatrix} 0 & 0 & -3 & 1 & 1 & 8 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & -6 & 2 & 2 & 16 \\ 0 & 0 & 3 & -1 & -1 & -8 \\ 0 & 0 & 9 & -3 & -3 & -24 \\ 0 & 0 & -3 & 1 & 1 & 8 \end{bmatrix}$$

a) Define c as the third column of A (which is its first nonzero column). Show that all columns of A are multiples of c, so $rank(A) = 1$.

b) Write $A = cr^\top$ for some vector r. How are the coefficients in $r^\top$ related to part (a)?

c) Give another example of a real rank 1 matrix and the corresponding c, r vectors for the decomposition $A = cr^\top$.

EXERCISE 16.4. *(Is the rank 1 decomposition unique?) Suppose A is a $m \times n$ rank 1 matrix with real entries, so $A = cr^\top$ for some nonzero $c \in \mathbb{R}^m, r \in \mathbb{R}^n$.*

a) Are the c and r vectors unique? That is, can we find other vectors p, q such that $A = pq^\top$?

b) Suppose we write a SVD for A as $A = \sigma uv^\top$, where $\sigma \in \mathbb{R}$, $\sigma \geq 0$, $u \in \mathbb{R}^m$, $v \in \mathbb{R}^n$, $||u|| = 1$, $||v|| = 1$ (using the standard Euclidean norm). Is σ unique? Are u, v unique? In particular, if we have $A = \tilde{\sigma}\tilde{u}\tilde{v}^\top$ for $\tilde{\sigma} \in \mathbb{R}$, $\tilde{\sigma} \geq 0$, $\tilde{u} \in \mathbb{R}^m, \tilde{v} \in \mathbb{R}^n$, $||\tilde{u}|| = 1$, $||\tilde{v}|| = 1$, do we necessarily have $\sigma = \tilde{\sigma}$? $u = \tilde{u}$? $v = \tilde{v}$? If so, why? If not, how many distinct ways can we write a SVD for A?

EXERCISE 16.5. *(Completing columns for an orthonormal matrix) Consider the matrix in* (16.9).

a) Find real numbers for the "?" entries of the 4×3 matrix U in (16.9) *such that U has orthonormal columns.*

b) Show there are exactly two ways to fill out the missing entries of the 3×3 matrix $V^\top$ in (16.9) *so that $V^\top$ has orthonormal rows. What are the two ways?*

c) Compute $U\Sigma V^\top$ to show the answer is indeed the A matrix of (16.9).

EXERCISE 16.6. *(Multiple choice – Computing singular values for bounds 1) Suppose*

$$A = \begin{bmatrix} 5 & 5 \\ 5 & 5 \\ 5 & 5 \end{bmatrix}$$

Find the tightest possible values of α, β *that ensure*

$$\alpha \, ||x|| \leq ||Ax|| \leq \beta \, ||x|| \quad \forall x \in \mathbb{R}^2$$

a) $(\alpha, \beta) = (0, 12.2474)$

b) $(\alpha, \beta) = (2.0790, 13.2166)$

c) $(\alpha, \beta) = (0.0035, 17.5556)$

d) $(\alpha, \beta) = (0, 18)$

EXERCISE 16.7. *(Multiple choice – Computing singular values for bounds 2) Suppose*

$$A = \begin{bmatrix} 4 & -2 \\ 0 & 2 \\ 7 & -9 \end{bmatrix}$$

Find the tightest possible values of α, β *that ensure*

$$\alpha \, ||x|| \leq ||Ax|| \leq \beta \, ||x|| \quad \forall x \in \mathbb{R}^2$$

a) $(\alpha, \beta) = (0, 15.4403)$

b) $(\alpha, \beta) = (1.9855, 7.2207)$

c) $(\alpha, \beta) = (-2.3303, 18)$

d) $(\alpha, \beta) = (2.2345, 12.2068)$

EXERCISE 16.8. *(Multiple choice – Computing singular values for bounds 3) Suppose*

$$A = \begin{bmatrix} 1 & 3 \\ 1 & 2 \\ 1 & -1 \\ 9 & 9 \end{bmatrix}$$

Find the tightest possible values of α, β that ensure

$$\alpha \, ||x|| \leq ||Ax|| \leq \beta \, ||x|| \quad \forall x \in \mathbb{R}^2$$

a) $(\alpha, \beta) = (2, 14)$

b) $(\alpha, \beta) = (2.0790, 13.2166)$

c) $(\alpha, \beta) = (0, 16.3333)$

d) $(\alpha, \beta) = (1.8888, 10.3042)$

EXERCISE 16.9. *(Multiple choice – Computing singular values for bounds 4) Suppose*

$$A = \begin{bmatrix} 9 & 1 & 7 & 1 \\ 0 & 2 & 2 & 3 \end{bmatrix}$$

Find the tightest possible values of α, β that ensure

$$\alpha \, ||x|| \leq ||Ax|| \leq \beta \, ||x|| \quad \forall x \in \mathbb{R}^4$$

a) $(\alpha, \beta) = (0, 8)$

b) $(\alpha, \beta) = (1.5684, 9.8909)$

c) $(\alpha, \beta) = (3.7339, 11.6214)$

d) $(\alpha, \beta) = (3/2, \sqrt{13})$

EXERCISE 16.10. *Consider the 2×2 matrix A given by*

$$A = \begin{bmatrix} \sqrt{1/2} & -\sqrt{1/2} \\ \sqrt{1/2} & \sqrt{1/2} \end{bmatrix} \begin{bmatrix} 8 & 0 \\ 0 & 3.2 \end{bmatrix} \begin{bmatrix} \sqrt{1/5} & 2/\sqrt{5} \\ \sqrt{4/5} & -1/\sqrt{5} \end{bmatrix}$$

a) What is $rank(A)$?

b) With $r = rank(A)$, write A in the SVD form $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$. Specify the singular values and the u_i, v_i vectors.

c) Find a vector x^* that solves

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 1, x \in \mathbb{R}^2 \end{aligned}$$

d) Find a vector x^* that solves

$$\begin{aligned} \textit{Minimize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 1, x \in \mathbb{R}^2 \end{aligned}$$

EXERCISE 16.11. *Consider the* 3×2 *matrix* A *given by*

$$A = \begin{bmatrix} \sqrt{1/7} & -\sqrt{1/2} \\ \sqrt{1/7} & \sqrt{1/2} \\ \sqrt{5/7} & 0 \end{bmatrix} \begin{bmatrix} 10 & 0 \\ 0 & 5 \end{bmatrix} \begin{bmatrix} \sqrt{1/5} & 2/\sqrt{5} \\ \sqrt{4/5} & -1/\sqrt{5} \end{bmatrix}$$

a) What is $rank(A)$*?*

b) With $r = rank(A)$*, write* A *in the SVD form* $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$*. Specify the singular values and the* u_i, v_i *vectors.*

c) Find a vector x^* *that solves*

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 1, x \in \mathbb{R}^2 \end{aligned}$$

d) Find a vector x^* *that solves*

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 100, x \in \mathbb{R}^2 \end{aligned}$$

EXERCISE 16.12. *Consider the* 3×2 *matrix* A *given by*

$$A = \begin{bmatrix} \sqrt{1/7} & -\sqrt{1/2} \\ \sqrt{1/7} & \sqrt{1/2} \\ \sqrt{5/7} & 0 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 0 & -50 \end{bmatrix} \begin{bmatrix} \sqrt{1/5} & 2/\sqrt{5} \\ \sqrt{4/5} & -1/\sqrt{5} \end{bmatrix}$$

a) What is $rank(A)$*?*

b) With $r = rank(A)$*, write* A *in the SVD form* $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$*. Specify the singular values and the* u_i, v_i *vectors.*

c) Find a vector x^* *that solves*

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 1, x \in \mathbb{R}^2 \end{aligned}$$

d) Find a vector x^* *that solves*

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 100, x \in \mathbb{R}^2 \end{aligned}$$

EXERCISE 16.13. *Consider the* 3×3 *matrix* A *given by*

$$A = \begin{bmatrix} \sqrt{1/7} & -\sqrt{1/2} & \cos(\sqrt{2}) \\ \sqrt{1/7} & \sqrt{1/2} & -\pi \\ \sqrt{5/7} & 0 & e^3/3 \end{bmatrix} \begin{bmatrix} 2 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{1/5} & 2/\sqrt{5} & 0 \\ \sqrt{4/5} & -1/\sqrt{5} & 0 \\ 888 & -1/e^{\pi} & 10^{10} \end{bmatrix}$$

a) What is $\mathrm{rank}(A)$*?*

b) With $r = \mathrm{rank}(A)$*, write* A *in the SVD form* $\sum_{i=1}^{r} \sigma_i u_i v_i^\top$*. Specify the singular values and the* u_i, v_i *vectors.*

c) Find a vector x^* *that solves*

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 1, x \in \mathbb{R}^3 \end{aligned}$$

d) Find a vector x^* *that solves*

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax|| \\ \textit{Subject to:} \quad & ||x|| = 100, x \in \mathbb{R}^3 \end{aligned}$$

EXERCISE 16.14. *Consider the* 4×4 *matrix* A *defined by*

$$A = \underbrace{10}_{\sigma_1} \underbrace{\begin{bmatrix} 1/2 \\ 1/2 \\ 1/2 \\ 1/2 \end{bmatrix}}_{u_1} \underbrace{\begin{bmatrix} \sqrt{1/6} & \sqrt{1/6} & \sqrt{2/6} & \sqrt{2/6} \end{bmatrix}}_{v_1^\top} + \underbrace{3}_{\sigma_2} \underbrace{\begin{bmatrix} -1/2 \\ 1/2 \\ 1/2 \\ -1/2 \end{bmatrix}}_{u_2} \underbrace{\begin{bmatrix} \sqrt{1/6} & -\sqrt{1/6} & \sqrt{2/6} & -\sqrt{2/6} \end{bmatrix}}_{v_2^\top}$$

a) Are $\{u_1, u_2\}$ *orthonormal? Are* $\{v_1, v_2\}$ *orthonormal?*

b) Compute Av_1*,* Av_2*,* $||Av_1||$*,* $||Av_2||$*.*

c) Find the tightest real numbers α, β *such that*

$$\alpha \, ||x|| \leq ||Ax|| \leq \beta \, ||x|| \qquad \forall x \in \mathbb{R}^4$$

Find a norm-1 vector $x^* \in \mathbb{R}^4$ *that achieves the upper bound.*

d) Give a representation $A = U\Sigma V^\top$ *for* Σ *nonnegative diagonal* 2×2*,* $U \sim 4 \times 2$ *with orthonormal columns,* $V \sim 4 \times 2$ *with orthonormal columns.*

e) Compute $A(v_1 + v_2)$ *and* $||A(v_1 + v_2)||^2$*.*

f) Compute $A(2v_1 - v_2)$ *and* $||A(2v_1 - v_2)||^2$*.*

EXERCISE 16.15. *Consider the $m \times n$ matrix A defined by*

$$A = \sum_{i=1}^{3} i u_i v_i^\top$$

where $\{u_1, u_2, u_3\}$ are orthonormal in $\mathbb{R}^m$ and $\{v_1, v_2, v_3\}$ are orthonormal in $\mathbb{R}^n$ (so $m \geq 3$, $n \geq 3$).

a) Find Av_1, Av_2, Av_3 and $||Av_1||^2, ||Av_2||^2, ||Av_3||^2$.

b) Find $A(v_1 + 7v_2 - 3v_3)$ and $||A(v_1 + 7v_2 - 3v_3)||^2$.

c) Suppose we complete $\{v_1, ..., v_3\}$ to an orthonormal basis $\{v_1, ..., v_n\}$ for $\mathbb{R}^n$. Fix $(\theta_1, ..., \theta_n) \in \mathbb{R}^n$. Find $A \sum_{i=1}^n \theta_i v_i$ and $\left\|A \sum_{i=1}^n \theta_i v_i\right\|^2$.

d) Find an $x \in \mathbb{R}^n$ that solves

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax||^2 \\ \textit{Subject to:} \quad & ||x||^2 = 1, x \in \mathbb{R}^n \end{aligned}$$

e) Find an $x \in \mathbb{R}^n$ that solves

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax||^2 \\ \textit{Subject to:} \quad & ||x||^2 = 2, x \in \mathbb{R}^n \end{aligned}$$

EXERCISE 16.16. *Let $\{u_1, u_2, ..., u_7\}$ be an orthonormal basis for $\mathbb{R}^7$. Consider the 7×7 matrix A defined by*

$$A = \sum_{i=1}^{3} (-2)^i u_i u_i^\top$$

a) Is A invertible? Is A symmetric? Is A positive semidefinite?

b) Compute $\text{rank}(A)$ and the singular values of A.

c) Compute $A(3u_1 - u_2 + u_6)$ and $||A(3u_1 - u_2 + u_6)||^2$.

d) Find an $x \in \mathbb{R}^7$ that solves

$$\begin{aligned} \textit{Maximize:} \quad & ||Ax||^2 \\ \textit{Subject to:} \quad & ||x||^2 = 5.2, x \in \mathbb{R}^7 \end{aligned}$$

e) Find an $x \in \mathbb{R}^7$ that solves

$$\begin{aligned} \textit{Minimize:} \quad & ||Ax||^2 \\ \textit{Subject to:} \quad & ||x||^2 = 5.2, x \in \mathbb{R}^7 \end{aligned}$$

EXERCISE 16.17. *Let $\{u_1, u_2, ..., u_7\}$ be an orthonormal basis for $\mathbb{R}^7$. Consider the 7×7 matrix A defined by*

$$A = \sum_{i=1}^{3} i^2 u_i u_i^\top$$

a) Is A invertible? Is A symmetric? Is A positive semidefinite?

b) Compute $rank(A)$ and the singular values of A.

c) Compute $(5u_1 + u_2 - 5u_4)^\top A(5u_1 + u_2 - 5u_4)$.

d) Fix $(\theta_1, ..., \theta_7) \in \mathbb{R}^7$ and define $x = \sum_{j=1}^{7} \theta_j u_j$. Compute $x^\top Ax$.

e) Find an $x \in \mathbb{R}^7$ that solves

$$\begin{aligned} &\textit{Maximize:} && x^\top Ax \\ &\textit{Subject to:} && ||x|| = 1, x \in \mathbb{R}^7 \end{aligned}$$

f) Find an $x \in \mathbb{R}^7$ that solves

$$\begin{aligned} &\textit{Maximize:} && x^\top Ax \\ &\textit{Subject to:} && ||x|| = 10, x \in \mathbb{R}^7 \end{aligned}$$

g) Find an $x \in \mathbb{R}^7$ that solves

$$\begin{aligned} &\textit{Minimize:} && x^\top Ax \\ &\textit{Subject to:} && ||x|| = 10, x \in \mathbb{R}^7 \end{aligned}$$

EXERCISE 16.18. *(Maximizing $|x^\top Ax|$) Solve the problem*

$$\begin{aligned} &\textit{Maximize:} && |x^\top Ax| \\ &\textit{Subject to:} && ||x|| = 3 \\ & && x \in \mathbb{R}^2 \end{aligned}$$

for

$$A = \begin{bmatrix} -7 & 1 \\ 1 & 2 \end{bmatrix}$$

EXERCISE 16.19. *Let $||A||$ denote the largest singular value of a matrix A. Design a 2×2 matrix A with real entries that satisfies $||A|| > 1$ but $\left|\left|A^2\right|\right| < 1$.*

EXERCISE 16.20. *(Simultaneous diagonalizable in general field – Part 1) Let $\mathbb{F}$ be a field. Let A*

and B be $n \times n$ matrices with entries in $\mathbb{F}$. We say that A and B are simultaneously diagonalizable *if there is a* common *invertible $n \times n$ matrix S such that*

$$A = S\Lambda_A S^{-1} \quad , \quad B = S\Lambda_B S^{-1}$$

for some $n \times n$ diagonal matrices Λ_A and Λ_B. Thus, not only are both A and B diagonalizable, but they have the same eigenvectors given by the columns of the matrix S.

a) Show that if A and B are simultaneously diagonalizable, then $AB = BA$ (so A and B commute).

b) Trivial examples: Fix $c \in \mathbb{F}$. Suppose A is a diagonalizable matrix. Show that if either $B = cI$, or $B = cA$, then A and B are simultaneously diagonalizable.

c) Nontrivial examples: Using any field you like, give an example of two simultaneously diagonalizable 2×2 matrices. Your example should avoid the trivial situations of part (b). Verify that your matrices indeed commute.

EXERCISE 16.21. *(Simultaneous diagonalizable in general field – Part 2 Electric Boogaloo) Let $\mathbb{F}$ be a field. Let A and B be $n \times n$ matrices with entries in $\mathbb{F}$. Suppose A is diagonalizable with n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$, and $AB = BA$. We want to show that A and B are simultaneously diagonalizable (recall the definition of simultaneously diagonalizable given in Exercise 16.20).*

a) Let (λ, v) be an eigenpair of A. Show that either $Bv = 0$, or Bv is an eigenvector of A with eigenvalue λ.

b) Suppose A has n distinct eigenvalues $\lambda_1, \ldots, \lambda_n$ with corresponding eigenvectors $\{v_1, \ldots, v_n\}$ that form a basis for $\mathbb{F}^n$. Use part (a) to show that for each $i \in \{1, \ldots, n\}$, v_i is an eigenvector of B (with some eigenvalue μ_i), and so A and B are simultaneously diagonalizable.

EXERCISE 16.22. *(Diagonalizable with general field) Let $\mathbb{F}$ be a field. Suppose matrix A has eigenvectors $\{v_1, \ldots, v_n\}$ that form a basis of $\mathbb{F}^n$, with corresponding (possibly repeated) eigenvalues $\lambda_1, \ldots, \lambda_n$. Suppose (μ, x) is an eigenpair of A.*

a) Must x be one of the n vectors in the set $\{v_1, \ldots, v_n\}$? Prove true, or provide a counterexample.

b) Must μ be one of the n scalars in the set $\{\lambda_1, \ldots, \lambda_n\}$? Prove true, or provide a counterexample.

EXERCISE 16.23. *Suppose $\mathbb{F} = \mathbb{R}$ and*

$$A = \begin{bmatrix} 5 & 4 & 0 \\ 2 & 1 & 1 \\ 3 & 6 & 7 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 5 & 4 & 0 \\ 2 & 1 & 1 \\ 3 & 6 & 7 \end{bmatrix}^{-1}$$

a) Is A diagonalizable? Is A invertible?

b) What are the eigenvalues of A?

c) Give an example of two different eigenvectors $x \in \mathbb{R}^3$ and $y \in \mathbb{R}^3$ that both have eigenvalue 2.

d) Give an example of two different eigenvectors $x \in \mathbb{R}^3$ and $y \in \mathbb{R}^3$ that both have eigenvalue 1.

EXERCISE 16.24. *Suppose $\mathbb{F} = \mathbb{R}$ and*

$$A = \begin{bmatrix} 0 & 2 & 4 \\ 1 & 1 & 7 \\ 1 & 2 & 7 \end{bmatrix} \begin{bmatrix} -4 & 0 & 0 \\ 0 & -4 & 0 \\ 0 & 0 & 3 \end{bmatrix} \begin{bmatrix} 0 & 2 & 4 \\ 1 & 1 & 7 \\ 1 & 2 & 7 \end{bmatrix}^{-1}$$

a) Is A diagonalizable? Is A invertible?

b) What are the eigenvalues of A?

c) Give an example of an eigenvector $x \in \mathbb{R}^3$ with eigenvalue -4 such that x is neither a scalar multiple of $(0, 1, 1)$ nor a scalar multiple of $(2, 1, 2)$.

d) Does A have an eigenvector $x \in \mathbb{R}^3$ with eigenvalue 3 such that x is not a scalar multiple of $(4, 7, 7)$?

EXERCISE 16.25. *Suppose $\mathbb{F} = \mathbb{R}$ and*

$$A = \begin{bmatrix} 7 & 4 & 0 & 9 \\ 2 & 1 & 1 & 0 \\ 3 & 6 & 7 & 1 \\ -1 & 2 & 0 & -3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 8 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} 7 & 4 & 0 & 9 \\ 2 & 1 & 1 & 0 \\ 3 & 6 & 7 & 1 \\ -1 & 2 & 0 & -3 \end{bmatrix}^{-1}$$

a) Is A diagonalizable? Is A invertible?

b) What are the eigenvalues of A?

c) For each eigenvalue of part (b), give two different eigenvectors with that eigenvalue.

EXERCISE 16.26. *Suppose $\mathbb{F} = \mathbb{R}$ and*

$$A = \begin{bmatrix} 7 & 4 & 0 & 9 \\ 2 & 1 & 1 & 0 \\ 3 & 6 & 7 & 1 \\ -1 & 2 & 0 & -3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix} \begin{bmatrix} 7 & 4 & 0 & 9 \\ 2 & 1 & 1 & 0 \\ 3 & 6 & 7 & 1 \\ -1 & 2 & 0 & -3 \end{bmatrix}^{-1}$$

a) Is A diagonalizable? Is A invertible?

b) What are the eigenvalues of A?

c) For each eigenvalue of part (b), give two different eigenvectors with that eigenvalue.

EXERCISE 16.27. *Suppose* $\mathbb{F} = \mathbb{R}$ *and*

$$S = \begin{bmatrix} 7 & 4 & 0 & 9 \\ 2 & 1 & 1 & 0 \\ 3 & 6 & 7 & 1 \\ -1 & 2 & 0 & -3 \end{bmatrix} \quad , \quad A = S \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} S^{-1}$$

Give an example of an eigenvector of A that is not a scalar multiple of one of the columns of S. What is its corresponding eigenvalue?

EXERCISE 16.28. *(Rank 1 matrices) Fix nonzero vectors* $c \in \mathbb{R}^m$ *and* $r \in \mathbb{R}^n$*. Define* $A = cr^\top$*.*

a) What is the size of A?

b) Show that $Ar \neq 0$*. Conclude that* $rank(A) \geq 1$*.*

c) Show that $Ax \in Span\{c\}$ *for all* $x \in \mathbb{R}^n$*. Conclude that* $rank(A) \leq 1$*.*

d) Write each of the following matrices in the form $cr^\top$ *for some nonzero vectors c, r, or briefly describe why it cannot be done:*

$$A = \begin{bmatrix} -3 & 3 & 6 \\ -4 & 4 & 8 \\ -5 & 5 & 10 \end{bmatrix}, \quad B = \begin{bmatrix} 24 & 6 & -9 & 3 \\ 32 & 8 & -12 & 4 \\ 40 & 10 & -15 & 5 \end{bmatrix}, \quad C = \begin{bmatrix} -3 & 3 & 7 \\ -4 & 4 & 9 \\ -5 & 5 & 11 \end{bmatrix}$$

e) For the matrices in part (d) that have the form $cr^\top$*, write their singular value decomposition* $\sigma uv^\top$*, where* $||u|| = ||v|| = 1$ *and* $\sigma > 0$*.*

f) For the matrices A, B, C of part (d), find the eigenvalues of $A^\top A$*,* $B^\top B$*,* $C^\top C$*.*

EXERCISE 16.29. *A is a real* 3×2 *matrix with rank 2, integer entries, and SVD* $A = \sum_{i=1}^{2} \sigma_i u_i v_i^\top$ *with*

$$\sigma_1 = 20.578484691240398 \quad , \quad \sigma_2 = 3.086416662147989$$

$$u_1 = \begin{bmatrix} -0.875547031658219 \\ -0.090184914098365 \\ -0.474640997621941 \end{bmatrix} \quad , \quad u_2 = \begin{bmatrix} -0.197659691663405 \\ -0.829576920000298 \\ 0.522238240743003 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} -0.582539121111295 \\ -0.812802665088446 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 0.812802665088446 \\ -0.582539121111295 \end{bmatrix}$$

where real numbers are approximated with 15 digits of precision.

a) Find the integer matrix A.

b) Write A in the SVD form $A = U\Sigma V^\top$*, where* Σ *is a* 2×2 *diagonal matrix of singular values of A.*

c) Find a vector $x^* \in \mathbb{R}^2$ that maximizes $||Ax||$ subject to $||x|| = 10$. What is the corresponding value of $||Ax^*||$?

d) Test three other x vectors in $\mathbb{R}^2$ that have $||x|| = 10$, none of which are multiples of your answer in part (c). Show that their $||Ax||$ values are strictly less than the value of part (c).

EXERCISE 16.30. *A is a real 3×2 matrix with rank 2, integer entries, and SVD $A = \sum_{i=1}^{2} \sigma_i u_i v_i^\top$ with*

$$\sigma_1 = 17.890809341203717 \quad , \quad \sigma_2 = 4.890699450661210$$

$$u_1 = \begin{bmatrix} -0.948412003482944 \\ -0.016018611148457 \\ -0.316635556667823 \end{bmatrix} \quad , \quad u_2 = \begin{bmatrix} 0.295724568909623 \\ -0.404719335546257 \\ -0.865302975135423 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} -0.989680316014025 \\ -0.143292958983961 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 0.143292958983961 \\ -0.989680316014025 \end{bmatrix}$$

where real numbers are approximated with 15 digits of precision.

a) Find the integer matrix A.

b) Write A in the SVD form $A = U\Sigma V^\top$, where Σ is a 2×2 diagonal matrix of singular values of A.

c) Find a vector $x^* \in \mathbb{R}^2$ that maximizes $||Ax||$ subject to $||x|| = 100$. What is the corresponding value of $||Ax^*||$?

d) Test three other x vectors in $\mathbb{R}^2$ that have $||x|| = 100$, none of which are multiples of your answer in part (c). Show that their $||Ax||$ values are strictly less than the value of part (c).

EXERCISE 16.31. *A is a real 3×2 matrix with rank 2, integer entries, and SVD $A = \sum_{i=1}^{2} \sigma_i u_i v_i^\top$ with*

$$\sigma_1 = 14.412010690562335 \quad , \quad \sigma_2 = 4.928889109638905$$

$$u_1 = \begin{bmatrix} -0.824736203288077 \\ -0.218481745187912 \\ -0.521608974237994 \end{bmatrix} \quad , \quad u_2 = \begin{bmatrix} 0.391335569964323 \\ -0.886340262175299 \\ -0.247502346105488 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} -0.536063511505855 \\ -0.844177654070523 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 0.844177654070523 \\ -0.536063511505855 \end{bmatrix}$$

where real numbers are approximated with 15 digits of precision.

a) Find the integer matrix A.

b) Write A in the SVD form $A = U\Sigma V^\top$, where Σ is a 2×2 diagonal matrix of singular values of A.

c) Find a vector $x^* \in \mathbb{R}^2$ that maximizes $||Ax||$ subject to $||x|| = 100$. What is the corresponding

value of $||Ax^*||$?

d) Compute $A(5v_1)$, Av_2, $A(5v_1 + v_2)$, $A(v_1 - v_2)$, $A(10v_1 + 8v_2)$.

e) Find a vector $x \in \mathbb{R}^2$ with norm 7 that satisfies $||Ax|| = 7\sigma_2$.

EXERCISE 16.32. *A is a real 3×2 matrix with rank 2, integer entries, and SVD $A = \sum_{i=1}^{2} \sigma_i u_i v_i^\top$ with*

$$\sigma_1 = 20.035668847618201 \quad , \quad \sigma_2 = 8.035668847618201$$

$$u_1 = \begin{bmatrix} -0.999364473010407 \\ -0.035646179102851 \\ 0 \end{bmatrix} \quad , \quad u_2 = \begin{bmatrix} -0.035646179102851 \\ 0.999364473010407 \\ 0 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} -0.999364473010407 \\ 0.035646179102851 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 0.035646179102851 \\ 0.999364473010407 \end{bmatrix}$$

where real numbers are approximated with 15 digits of precision.

a) Find the integer matrix A.

b) Write A in the SVD form $A = U\Sigma V^\top$, where Σ is a 2×2 diagonal matrix of singular values of A.

c) Find a vector $x^* \in \mathbb{R}^2$ that maximizes $||Ax||$ subject to $||x|| = 10$. What is the corresponding value of $||Ax^*||$?

d) Compute $A(5v_1)$, Av_2, $A(5v_1 + v_2)$, $A(v_1 - v_2)$, $A(10v_1 + 8v_2)$.

e) Find a vector $x \in \mathbb{R}^2$ with norm 7 that satisfies $||Ax|| = 7\sigma_2$.

EXERCISE 16.33. *(Singular values: True or false) Let A be a real $m \times n$ matrix with rank $r \geq 1$.*

a) True or false: $rank(A) = rank(A^\top)$.

b) True or false: A has r nonzero singular values.

c) True or false: $A^\top$ has r nonzero singular values.

d) True or false: The nonzero singular values of A are the same as the nonzero singular values of $A^\top$.

EXERCISE 16.34. *(Computing SVD for A and $A^\top$) Define*

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

a) Define $r = rank(A)$. Find r.

b) Find an SVD for A. Find both SVD forms $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ and $A = U\Sigma V^\top$.

c) Find an SVD for $A^\top$.

EXERCISE 16.35. *(Computing SVD for A and $A^\top$) Define*

$$A = \begin{bmatrix} 3 & 2 \\ 1 & 1 \\ -1 & 1 \end{bmatrix}$$

a) Define $r = rank(A)$. Find r.

b) Find an SVD for A. Find both SVD forms $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ and $A = U\Sigma V^\top$.

c) Find an SVD for $A^\top$.

EXERCISE 16.36. *(Computing SVD for A and $A^\top$) Define*

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

a) Define $r = rank(A)$. Find r.

b) Find an SVD for A. Find both SVD forms $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ and $A = U\Sigma V^\top$.

c) Find an SVD for $A^\top$.

EXERCISE 16.37. *(Computing SVD for A and $A^\top$) Define*

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

a) Define $r = rank(A)$. Find r.

b) Find an SVD for A. Find both SVD forms $A = \sum_{i=1}^{r} \sigma_i u_i v_i^\top$ and $A = U\Sigma V^\top$.

c) Find an SVD for $A^\top$.

EXERCISE 16.38. *A is a real 3×2 matrix with rank 2, integer entries, and SVD $A = \sum_{i=1}^{2} \sigma_i u_i v_i^\top$ with*

$$\sigma_1 = 3.784779428227853 \quad , \quad \sigma_2 = 1.294389693895637$$

$$u_1 = \begin{bmatrix} 0.521608974237994 \\ -0.218481745187912 \\ 0.824736203288077 \end{bmatrix} \quad , \quad u_2 = \begin{bmatrix} -0.247502346105488 \\ 0.886340262175299 \\ 0.391335569964323 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} 0.160182243006967 \\ 0.987087457637497 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 0.987087457637497 \\ -0.160182243006967 \end{bmatrix}$$

where real numbers are approximated with 15 digits of precision.

a) Find an SVD for $A^\top$.

b) Find the (integer) matrix A.

c) Write A in the SVD form $A = U\Sigma V^\top$, where Σ is a 2×2 diagonal matrix of singular values of A.

d) Write $A^\top$ in the SVD form $A^\top = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$, where $\tilde{\Sigma}$ is a 2×2 diagonal matrix of singular values of $A^\top$.

EXERCISE 16.39. *A is a real 2×3 matrix with rank 2, integer entries, and SVD $A = \sum_{i=1}^{2} \sigma_i u_i v_i^\top$ with*

$$\sigma_1 = 5.843870789583452 \quad , \quad \sigma_2 = 0.921506481069606$$

$$u_1 = \begin{bmatrix} -0.935605022678575 \\ -0.353048497431475 \end{bmatrix} \quad , \quad u_2 = \begin{bmatrix} -0.353048497431475 \\ 0.935605022678575 \end{bmatrix}$$

$$v_1 = \begin{bmatrix} -0.921328055003011 \\ -0.320200448081866 \\ -0.220513691440105 \end{bmatrix} \quad , \quad v_2 = \begin{bmatrix} 0.114993828450101 \\ -0.766242027992439 \\ 0.632178435219380 \end{bmatrix}$$

where real numbers are approximated with 15 digits of precision.

a) Find an SVD for $A^\top$.

b) Find the (integer) matrix A.

c) Write A in the SVD form $A = U\Sigma V^\top$, where Σ is a 2×2 diagonal matrix of singular values of A.

d) Write $A^\top$ in the SVD form $A^\top = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$, where $\tilde{\Sigma}$ is a 2×2 diagonal matrix of singular values of $A^\top$.

EXERCISE 16.40. *(SVD for powers of A) Let A be a real $n \times n$ matrix. Define σ as its largest singular value, so $||Ax|| \leq \sigma ||x||$ for all $x \in \mathbb{R}^n$.*

a) Show that $\left||A^2 x|\right| \leq \sigma^2 ||x||$ for all $x \in \mathbb{R}^n$.

b) Show that $\left||A^k x|\right| \leq \sigma^k ||x||$ for all $x \in \mathbb{R}^n$ and all $k \in \{1, 2, 3, \ldots\}$.

c) Compute the largest singular values of A, A^2, and A^3 for

$$A = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix}$$

EXERCISE 16.41. *(SVD of a matrix with one row) Define the 1×4 matrix*

$$A = [1\,1\,1\,1]$$

a) Find $rank(A)$.

b) Find the SVD of A.

c) Find the SVD of $A^\top$.

d) Find a vector $x^* \in \mathbb{R}^4$ that maximizes $||Ax||$ subject to $||x|| = 1$.

EXERCISE 16.42. *(SVD of a matrix with three identical rows) Define the 3×4 matrix*

$$A = \begin{bmatrix} 1 & 2 & 9 & 0 \\ 1 & 2 & 9 & 0 \\ 1 & 2 & 9 & 0 \end{bmatrix}$$

a) Find $rank(A)$.

b) Find the SVD of A.

c) Find the SVD of $A^\top$.

d) Find a vector $x^* \in \mathbb{R}^4$ that maximizes $||Ax||$ subject to $||x|| = 1$.

EXERCISE 16.43. *(SVD of a matrix with identical columns) Let A be a real $m \times n$ matrix with n identical columns. Let $y \in \mathbb{R}^m$ be the first column. Assume $y \neq 0$.*

a) What is the rank of A?

b) What is the SVD of A? What is the largest singular value of A?

c) What is the SVD of $A^\top$? What is the largest singular value of $A^\top$?

EXERCISE 16.44. *(Norm of a matrix) Let A be a real $m \times n$ matrix. Define $||A||$ as the largest singular value of A. This is called the* 2-norm *or the* spectral norm *of the matrix. We know that*

$$||A|| = \sup_{x \in \mathbb{R}^n : ||x|| = 1} \{||Ax||\}$$

Let V denote the vector space of all real $m \times n$ matrices. This exercise shows the spectral norm satisfies the requirements for a norm on the vector space V.

a) Show that $||A|| \geq 0$ for $A \in V$, with equality if and only if A is the all-zero matrix.

b) Show that $||cA|| = |c| \cdot ||A||$ for all $c \in \mathbb{R}$ and all $A \in V$.

c) Show the triangle inequality: $||A + B|| \leq ||A|| + ||B||$ for all $A, B \in V$.

d) Compute $||A||$, $||B||$, and $||A + B||$ for

$$A = \begin{bmatrix} 1 & 2 \\ 0 & 3 \\ -1 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 3 & 0 \\ 1 & 0 \\ 1 & 1 \end{bmatrix}$$

EXERCISE 16.45. *(Norm of a matrix) Let A be a real $m \times n$ matrix. Define the spectral norm of the matrix A, denoted $||A||$, as in Exercise 16.44.*

a) Find $||A||$ *when*

$$A = \begin{bmatrix} 2 & 0 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}$$

b) Find real 3×2 *matrices* A, B *such that* $||A|| = ||B|| = 1$ *and* $||A + B|| = 2$*. (Hint: Keep it as simple as possible.)*

c) What is $||A||$ *when* A *is a* $m \times n$ *matrix with identical rows?*

d) What is $||A||$ *when* A *is a* $m \times n$ *matrix with identical columns?*

EXERCISE 16.46. *(*$||A|| = \left|\left|A^\top\right|\right|$*) Let* A *be a real* $m \times n$ *matrix. Define the spectral norm of the matrix* A*, denoted* $||A||$*, as in Exercise 16.44. Explain why* $||A|| = \left|\left|A^\top\right|\right|$*. Give an example.*

PART V.

Appendices

APPENDIX A.

Worksheet 1: Speed

Suppose A is a $m \times n$ matrix with entries in some field $\mathbb{F}$.

a) Let $y \in Col(A)$. Then what?

b) Let $x \in Null(A)$. Then what?

c) Let $z \in Row(A)$. Then what?

d) Let $w \in Null(A^\top)$. Then what?

e) What is the size of $A^\top A$? Of $AA^\top$?

f) What relationship between m, n ensures $Null(A) \neq \{0\}$?

g) What relationship between m, n ensures $Col(A) \neq \mathbb{F}^m$?

A.1. Speed 2

Suppose $\{x_1, ..., x_n\}$ are vectors in some vector space V over a field $\mathbb{F}$.

a) Suppose $y \in Span\{x_1, ..., x_n\}$. Then what?

b) What is the definition of $\{x_1, ..., x_n\}$ being linearly independent?

c) What does it mean for $\{x_1, ..., x_n\}$ to be a basis for V?

A.2. Speed with inner products

Suppose $w, u, v, x_1, ..., x_n$ are vectors in some real inner product space V.

a) Let $y \in (Span\{x_1, ..., x_n\})^\perp$. Then what?

b) Suppose $\langle w, w \rangle = 0$. Then what?

c) $||u + v||^2 =$

A.3. Speed with $\mathbb{R}^n$ and $\mathbb{R}^m$

Let A, B, C be matrices with real entries. Assume they have appropriate size when they are being multiplied. Use the standard inner product $\langle x, y\rangle = x^\top y$ and $||x||^2 = x^\top x$.

a) $(ABC)^\top =$

b) $(AB^\top C)^\top =$

c) $||Ax||^2 =$

d) $||ABx||^2 =$

e) $Null(A)^\perp =$

f) $Col(A)^\perp =$

A.4. Solving $Ax = b$ via transposes

Now that we are using transposes extensively for least squares problems, it will be fun to play around with the following new lemmas. For example, try proving one or more parts, or finding examples in interesting special cases and corner cases.

LEMMA A.1. *Suppose A is a $m \times n$ matrix with entries in some field $\mathbb{F}$. Fix $b \in \mathbb{F}^m$. We want to solve*

$$Ax = b \tag{A.1}$$

a) If $AA^\top$ is invertible then a (possibly nonunique) solution to (A.1) *is*

$$x^* = A^\top (AA^\top)^{-1} b$$

b) If $A^\top A$ is invertible, and if (A.1) *has a solution, then the solution is unique and is given by*

$$x^* = (A^\top A)^{-1} A^\top b$$

c) If neither $AA^\top$ nor $A^\top A$ is invertible, then (A.1) *may or may not have a solution.*

LEMMA A.2. *If A is a $m \times n$ matrix with real entries then*

a) $Null(A^\top A) = Null(A)$

b) $Null(AA^\top) = Null(A^\top)$

c) The properties in (a) and (b) require inner products: They do not necessarily hold when A has entries in a finite field $\mathbb{F}$.

APPENDIX B.

Worksheet 2: Pseudoinverse

This worksheet builds strength in linear algebra proofs. It also reviews matrix multiplication, linear independence, dimension, nullspace, invertible matrices, and transposes. It does this by considering two new definitions: *left pseudoinverse* and *right pseudoinverse*.

Let A be a $m \times n$ matrix with entries in $\mathbb{R}$.

- Definition: A matrix B is called a *left pseudoinverse of* A if $BA = I$.
- Definition: A matrix B is called a *right pseudoinverse of* A if $AB = I$.

We want to show that:

1. A has a left pseudoinverse if and only if its columns are linearly independent (so A is either square or tall).
2. A has a right pseudoinverse if and only if its rows are linearly independent (so A is either square or wide).
3. A has a *unique* left pseudoinverse if and only if A is square and invertible.
4. A has a *unique* right pseudoinverse if and only if A is square and invertible.

B.1. Null space

This section considers the null space of a matrix D.

a) Suppose D is a $m \times n$ matrix with entries in $\mathbb{R}$. Prove that $Null(D) = Null(D^\top D)$.

b) Suppose D is a $m \times n$ matrix with entries in $\mathbb{C}$. Prove that $Null(D) = Null(D^H D)$.

B.2. One direction

Recall that A has size $m \times n$. Suppose A has a left pseudoinverse B, so $BA = I$.

a) Find the size of B and the size of I (this reviews matrix multiplication).

b) Prove that the columns of A are linearly independent (this reviews linear independence proofs).

c) Conclude that A is either square or tall (this reviews dimension concepts).

B.3. Other direction

Recall that A has size $m \times n$ and its entries are real. Suppose the columns of A are linearly independent.

a) Find the size of $A^\top A$.

b) Show that $Null(A^\top A) = Null(A) = \{0\}$. Conclude that $A^\top A$ is invertible.

c) Use the fact that $A^\top A$ is invertible to construct an explicit left-pseudoinverse of A.

B.4. Right pseudoinverse

Use the fact that A has a left pseudoinverse if and only if its columns are linearly independent to prove that A has a right pseudoinverse if and only if its rows are linearly independent. Assuming the rows of A are linearly independent, construct an explicit right-pseudoinverse of A.

B.5. Example

Consider the following tall matrix:

$$A = \begin{bmatrix} 2 & -1 \\ 0 & 1 \\ 3 & 0 \end{bmatrix}$$

a) Find a left-pseudoinverse of A.

b) Find a different left-pseudoinverse of A. Hint: Add something to your answer of part (a); look at the left-nullspace of A (which is the nullspace of $A^\top$).

c) Find two distinct right-pseudoinverses of $A^\top$.

B.6. Tall matrices A

This section considers general tall matrices with real values.

a) Show that if a real-valued matrix A is tall and has linearly independent columns, then it has infinitely many left pseudoinverses.

b) Show that if a real-valued matrix A is square and has linearly independent columns, then it has one unique left pseudoinverse.

APPENDIX C.

Affine sets

Let $\mathbb{F}$ be a field. Let V be a vector space over $\mathbb{F}$. Elements of V are *vectors* and elements of $\mathbb{F}$ are *scalars*. The vector space V is allowed to be infinite dimensional.

DEFINITION C.1. *Let D be a nonempty subset of V. An* affine combination of points in D *is a vector of the form*

$$\sum_{i=1}^{k} a_i v_i$$

for some positive integer k, some vectors $v_1, \ldots, v_k \in V$, and some scalars $a_1, \ldots, a_k \in \mathbb{F}$ that satisfy $\sum_{i=1}^{k} a_i = 1$.

Therefore, an affine combination is a special type of linear combination where the coefficients sum to 1.

DEFINITION C.2. *A set $X \subseteq V$ is an* affine set *if it contains all of its affine combinations. That is, X is affine if for all positive integers k, all vectors $v_1, \ldots, v_k \in X$ and all scalars $a_1, \ldots, a_k \in \mathbb{F}$ that satisfy $\sum_{i=1}^{k} a_i = 1$, we have*

$$\sum_{i=1}^{k} a_i v_i \in X$$

By convention, the empty set is considered to be affine.

LEMMA C.1. *Let X be a nonempty subset of V. The set X is affine if and only if*

$$X = x^* + S$$

for some $x^* \in X$ and some subspace $S \subseteq V$. Here we use the Minkowski definition of $x^* + S$:

$$x^* + S = \{x^* + s : s \in S\}$$

Lemma C.1 ensures every subspace of V is itself an affine set (just use $x^* = 0$). If $\mathbb{F}$ is a finite field with $|\mathbb{F}| = f$, where $f \geq 2$ is an integer, then every finite dimensional subspace $S \subseteq V$ has size $|S| = f^k$ for some $k \in \{0, 1, 2, \ldots\}$. Thus, all nonempty affine sets X with a finite number of elements must have size $|X| = f^k$ for some nonnegative integer k.

PROOF. (Lemma C.1) First suppose X has the form $X = x^* + S$ for some $x^* \in X$ and some subspace $S \subseteq V$. It is a straightforward exercise to show that X is affine. [To start, fix $k \in \mathbb{N}$, let $v_1, \ldots, v_k$ be vectors in X, let $a_1, \ldots, a_k$ be scalars that sum to 1. Show that $\sum_{i=1}^k a_i v_i \in X$. Recall that a subspace S must contain all linear combinations of points in S.]

Now suppose X is affine. Since X has been assumed to be nonempty, choose x^* as any point in X. Define the set

$$S = X - x^* = \{x - x^* : x \in X\}$$

It can be shown that $X = x^* + S$. It suffices to show S is a subspace of V. Since $x^* \in X$, we know $x^* - x^* \in S$ and so $0 \in S$. Now suppose $s_1, s_2 \in S$ and $a_1, a_2 \in \mathbb{F}$. We want to show $a_1 s_1 + a_2 s_2 \in S$. Since $s_1, s_2 \in S$ we know

$$s_1 = x_1 - x^*$$
$$s_2 = x_2 - x^*$$

for some $x_1, x_2 \in X$. Then

$$\begin{aligned} a_1 s_1 + a_2 s_2 &= a_1(x_1 - x^*) + a_2(x_2 - x^*) \\ &= (a_1 x_1 - a_1 x^* + a_2 x_2 - a_2 x^* + 1 \cdot x^*) - x^* \end{aligned} \tag{C.1}$$

Define

$$y = (a_1 x_1 - a_1 x^* + a_2 x_2 - a_2 x^* + 1 \cdot x^*)$$

Note that y is an affine combination of points in X because $x_1, x^*, x_2, x^*, x^* \in X$ and $a_1 - a_1 + a_2 - a_2 + 1 = 1$. Since X is an affine set, it follows that $y \in X$. Substituting the definition of y into (C.1) gives

$$a_1 s_1 + a_2 s_2 = y - x^*$$

Since $y \in X$, we know $y - x^* \in S$, and so $a_1 s_1 + a_2 s_2 \in S$. Thus, S is indeed a subspace. □

There is nothing special about x^* in the above lemma. If $X = x^* + S$ for some $x^* \in X$ and some subspace S, it can be shown that

$$X = x + S \quad \forall x \in X$$

So any point of X can be used in the place of x^*, and the subspace S remains the same (the proof is left as an exercise). However, the subspace S associated with the affine set is fundamental: If

$$A = x_1 + S_1 = x_2 + S_2$$

for some vectors $x_1, x_2 \in A$ and some subspaces S_1, S_2, then we must have $S_1 = S_2$ (the proof is left as an exercise). The *dimension* of the affine set A is defined as $\dim(S)$ where S is the fundamental subspace associated with A. It can be shown that an affine set is a subspace if and only if it contains the vector 0.

C.1. Three points

Continue to assume V is a (possibly infinite dimensional) vector space over a general field $\mathbb{F}$. The definition of an affine set X requires X to contain all affine combinations of k points in X, for all positive integers k. It suffices to check that the condition holds only for affine combinations of three points of X.

LEMMA C.2. *(Affine combinations of three points) A set $X \subseteq V$ is affine if and only if for all vectors $v_1, v_2, v_3 \in V$ and all scalars $a_1, a_2, a_3 \in \mathbb{F}$ that satisfy $\sum_{i=1}^{3} a_i = 1$ we have*

$$\sum_{i=1}^{3} a_i v_i \in X$$

PROOF. First suppose X is an affine set. Then it contains all affine combinations of any positive integer number of vectors in X, so it clearly contains all affine combinations of three points in X and we are done.

Now suppose X contains all affine combinations of three points of X. Then it also contains all affine combinations of two points or one point in X. Indeed:

- One point: If $x \in X$ then $1 \cdot x \in X$.
- Two points: If $x_1, x_2 \in X$ and a_1, a_2 are scalars that satisfy $a_1 + a_2 = 1$ then

  $$a_1 x_1 + a_2 x_2 = a_1 x_1 + a_2 x_2 + 0 x_2$$

  The vector $a_1 x_1 + a_2 x_2 + 0 x_2$ is an affine combination of the three points x_1, x_2, x_2 in X, so we know $a_1 x_1 + a_2 x_2 + 0 x_2 \in X$.

We now use induction: Suppose X contains all affine combinations of k or fewer points in X, for some integer $k \geq 3$ (we already know this holds for $k = 3$). We prove it holds for $k+1$. Let $v_1, \ldots, v_{k+1}$ be vectors in X and let $a_1, \ldots, a_{k+1}$ be scalars that sum to 1. We want to show $\sum_{i=1}^{k+1} a_i v_i \in X$.

We first claim there is a subset $I \subseteq \{1, \ldots, k+1\}$ of size $|I|$ that satisfies $2 \leq |I| \leq k$ and $\sum_{i \in I} a_i \neq 0$. This is because

$$1 = (a_1 + a_2) + (a_3 + \ldots + a_{k+1})$$

so it is impossible to have both $(a_1 + a_2) = 0$ and $(a_3 + \ldots + a_{k+1}) = 0$. So we can take our subset I to be either $\{1, 2\}$ or $\{3, \ldots, k+1\}$. [Since $k \geq 3$, the set $\{3, \ldots, k+1\}$ contains at least two indices.]

Without loss of generality, assume our set I has the form $I = \{1, 2, \ldots, m\}$ for some integer m that satisfies $2 \leq m \leq k$ (if needed, we can just reorder the indices to make this true). So $\sum_{i=1}^{m} a_i \neq 0$. Then $(\sum_{i=1}^{m} a_i)^{-1}$ exists. Define a scalar θ and a vector y by

$$\theta = \left(\sum_{i=1}^{m} a_i \right)^{-1}$$

$$y = \sum_{i=1}^{m} (\theta a_i) v_i$$

Note that $\theta a_1, \ldots, \theta a_m$ are scalars that sum to 1, so y is an affine combination of m points of X, where $2 \leq m \leq k$. By our induction assumption, we know $y \in X$. Then

$$\begin{aligned} \sum_{i=1}^{k+1} a_i v_i &= \sum_{i=1}^{m} a_i v_i + \sum_{i=m+1}^{k+1} a_i v_i \\ &= \theta^{-1} y + \sum_{i=m+1}^{k+1} a_i v_i \end{aligned} \tag{C.2}$$

However,

$$\theta^{-1} + \sum_{i=m+1}^{k+1} a_i = \sum_{i=1}^{m} a_i + \sum_{i=m+1}^{k+1} a_i = 1$$

Thus, the right-hand-side of (C.2) is an affine combination of vectors $y, v_{m+1}, \ldots, v_{k+1} \in X$. Since $m \geq 2$, these vectors number k or fewer, so our induction assumption ensures

$$\theta^{-1} y + \sum_{i=m+1}^{k+1} a_i v_i \in X$$

By (C.2), it follows that $\sum_{i=1}^{k+1} a_i v_i \in X$ and we are done. □

C.2. Fields for which $1 + 1 \neq 0$

Let $\mathbb{F}$ be a field. The field $\mathbb{F}$ includes scalars 0 and 1. When can we claim $1 + 1 \neq 0$?

- The case $1+1 \neq 0$: Fields for which $1+1 \neq 0$ include $\mathbb{R}$ (the real numbers), $\mathbb{C}$ (the complex numbers), $\mathbb{Q}$ (the rational numbers), and the finite fields of size p^k for p an odd prime and $k \in \mathbb{N}$.
- The case $1 + 1 = 0$: Fields for which $1 + 1 = 0$ include finite fields of size 2^k for $k \in \mathbb{N}$.

C.3. Two points

Here we assume $\mathbb{F}$ is a field under which $1 + 1 \neq 0$. Suppose V is a (possibly infinite dimensional) vector space over this field $\mathbb{F}$. Then a set $X \subseteq V$ can be determined to be affine just by checking all affine combinations of two points of X.

LEMMA C.3. *(Affine combinations of two points) If the field $\mathbb{F}$ satisfies $1 + 1 \neq 0$, then a set $X \subseteq V$ is affine if and only if for all vectors $x, y \in X$ and all scalars $a \in \mathbb{F}$ we have*

$$ax + (1 - a)y \in X$$

PROOF. First suppose X is an affine set. Then it contains all affine combinations of any positive integer number of vectors in X, so it clearly contains all affine combinations of two points in X and we are done.

Conversely, suppose X contains all affine combinations of two points in X. We want to show the set X is affine. It suffices to verify that all affine combinations of three points of X are in X. Let v_1, v_2, v_3 be vectors in X. Let a_1, a_2, a_3 be scalars in $\mathbb{F}$ that satisfy

$$a_1 + a_2 + a_3 = 1$$

We want to show that $a_1v_1 + a_2v_2 + a_3v_3 \in X$.

It is impossible for $a_1 = a_2 = a_3 = 1$ because then $1 + 1 + 1 = 1$, meaning $1 + 1 = 0$, which would contradict our assumption $1 + 1 \neq 0$. Thus, at least one of the three scalars is not 1. Without loss of generality, assume $a_3 \neq 1$ (if this does not hold, just reorder the three indices to make it true). Then $a_1 + a_2 \neq 0$, so $(a_1 + a_2)^{-1}$ exists. Define

$$y = a_1(a_1 + a_2)^{-1}v_1 + a_2(a_1 + a_2)^{-1}v_2$$

Vector y is an affine combination of two points in X, so we know $y \in X$. Then

$$\begin{aligned} &a_1v_1 + a_2v_2 + a_3v_3 \\ &= (a_1 + a_2)\left[a_1(a_1 + a_2)^{-1}v_1 + a_2(a_1 + a_2)^{-1}v_2\right] + a_3v_3 \end{aligned}$$

$$= (a_1 + a_2)y + a_3 v_3$$

But $(a_1 + a_2)y + a_3 v_3$ is an affine combination of two points in X, so we know it is in X. □

C.4. Counter-example when $1 + 1 = 0$

Consider the binary field $\mathbb{F}_2 = \{0, 1\}$ with mod 2 arithmetic (so $1+1 = 0$). Let V be any vector space over the field $\mathbb{F}_2$. Then *all* subsets $X \subseteq V$ satisfy the property that for all $x, y \in X$ and all $a \in \mathbb{F}_2$, we have

$$ax + (1 - a)y \in X$$

Indeed, if X is the empty set then this holds true vacuously. Now let X be any nonempty subset of V. Let x, y be vectors in X. Fix $a \in \mathbb{F}_2$. Then

- If $a = 0$ then $ax + (1 - a)y = y$.
- If $a = 1$ then $ax + (1 - a)y = x$.

Thus, in the binary field $\mathbb{F}_2 = \{0, 1\}$, taking an affine combination of two points $x, y \in X$ gives one of those same points back! Affine combinations of two points of X cannot escape the set X.

This does not mean that all subsets of V are affine. Indeed, suppose the vector space V contains at least three distinct points v_1, v_2, v_3. Define $X = \{v_1, v_2, v_3\}$. The set X cannot be affine because it is nonempty and has a finite number of elements, but its number of elements is not a power of 2. So X cannot be written in the form $X = x^* + S$ for some subspace $S \subseteq V$.

For a more concrete example, define

$$V = \mathbb{F}_2^2 = \{(0, 0), (0, 1), (1, 0), (1, 1)\}$$

Define the set $X \subseteq V$ by

$$X = \{(0, 0), (0, 1), (1, 0)\}$$

Then X is not an affine set because $1 + 1 + 1 = 1$ but

$$1(0, 0) + 1(0, 1) + 1(1, 0) = (1, 1) \notin X$$

On the other hand, X satisfies the property that for all $x, y \in X$ and all $a \in \mathbb{F}_2$ we have

$$ax + (1 - a)y \in X$$

This counter-example does not contradict Lemma C.3 because the field $\mathbb{F}_2$ has $1+1=0$. When working with vector spaces over the binary field $\mathbb{F}_2$, to show a subset X is affine we need to show X contains all affine combinations of three of its points (rather than just two of its points). Or, we can just show $X = x^* + S$ for some $x^* \in X$ and some subspace $S \subseteq V$.

C.5. Intersections

Let V be a vector space over some field $\mathbb{F}$. The intersection of an arbitrary (possibly uncountably infinite) number of affine subspaces of V can also be shown to be an affine subspace of V. Of course, the intersection might be the empty set. This is one reason for our convention that the empty set is considered to be affine.

To formally prove this intersection fact about affine sets, let I be an arbitrary (possibly uncountably infinite) set that is used for indexing. For each $i \in I$ let X_i be an affine subset of V. We want to show $\cap_{i \in I} X_i$ is also affine. If $\cap_{i \in I} X_i = \phi$ then we are done because the empty set is affine (here ϕ denotes the empty set). If $\cap_{i \in I} X_i \neq \phi$, fix $k \in \mathbb{N}$ and let $v_1, \ldots, v_k$ be any vectors in $\cap_{i \in I} X_i$. Let $a_1, \ldots, a_k$ be any scalars that sum to 1. We want to show $\sum_{c=1}^{k} a_c v_c \in \cap_{i \in I} X_i$.

Fix $j \in I$. The vectors $v_1, \ldots, v_k$ are all in $\cap_{i \in I} X_i$, so they must also be in set X_j. Since X_j is affine we must have

$$\sum_{c=1}^{k} a_c v_c \in X_j$$

This holds for all $j \in I$. Thus

$$\sum_{c=1}^{k} a_c v_c \in \cap_{i \in I} X_i$$

and so $\cap_{i \in I} X_i$ is affine.

APPENDIX D.

Geometry of $\mathbb{R}^n$

This appendix treats $\mathbb{R}^n$. All matrices and vectors are assumed to have real entries. We consider the geometry of $Ax = b$. We also consider rotations in $\mathbb{R}^2$ and $\mathbb{R}^3$. This is useful for applications in computer graphics, video gaming, robotics, and aircraft control.

D.1. Row and column perspectives for $Ax = b$

Let A be a real $m \times n$ matrix. Fix $b \in \mathbb{R}^m$. The problem of finding a vector $x \in \mathbb{R}^n$ to solve $Ax = b$ can be viewed from a *column perspective* and a *row perspective*.

D.1.1. Column perspective

View A according to its columns

$$A = \begin{bmatrix} | & | & \cdots & | \\ c_1 & c_2 & \cdots & c_n \\ | & | & \cdots & | \end{bmatrix}$$

Each column c_j is a vector in $\mathbb{R}^m$. Solving $Ax = b$ amounts to finding scalar weights $x_1, \ldots, x_n$ to linearly combine the columns to form b:

$$x_1 c_1 + x_2 c_2 + \cdots + x_n c_n = b$$

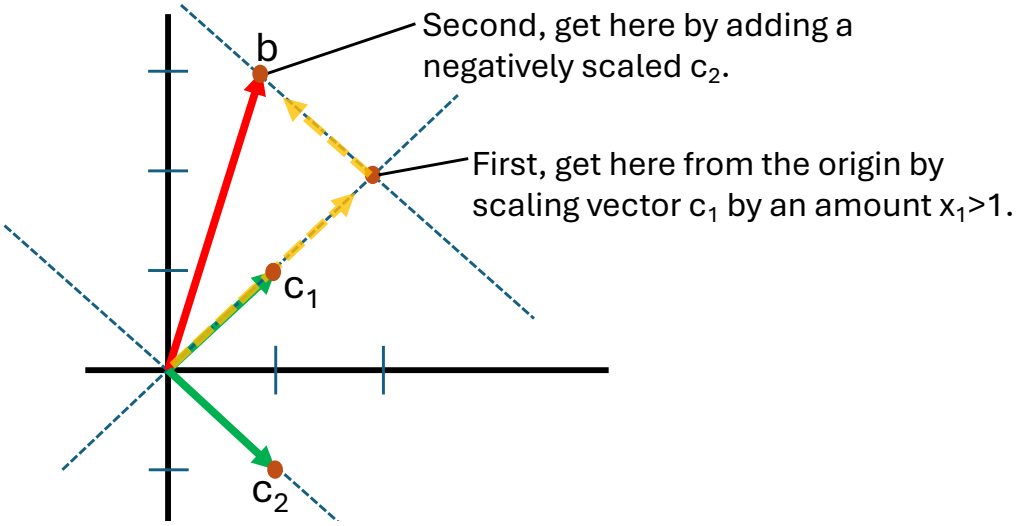


FIGURE D.1. Column perspective: Solving $Ax = b$ by linearly combining the two columns of A for the example (D.1).

For example, consider

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \quad , \quad b = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

Solving $Ax = b$ means finding real numbers x_1, x_2 that satisfy

(D.1)
$$x_1 \underbrace{\begin{bmatrix} 1 \\ 1 \end{bmatrix}}_{c_1} + x_2 \underbrace{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}_{c_2} = \underbrace{\begin{bmatrix} 1 \\ 3 \end{bmatrix}}_{b}$$

To visualize how this is possible:

1. Draw the vectors $c_1 = (1, 1)$, $c_2 = (1, -1)$, $b = (1, 3)$ on the 2-dimensional plane (see Fig. D.1).
2. Extend the vectors c_1 and c_2 to two infinite lines, each line passing through the origin and including the corresponding c_i vector.
3. Draw a line starting from point b that is parallel to the line of c_2. You will find this line intersects the line of c_1 at exactly one point.

As shown in Fig. D.1, the point of intersection suggests that to get $x_1c_1 + x_2c_2 = b$, we need $x_1 > 1$ and $x_2 < 0$. It is easy to solve $Ax = b$ in this example to get the unique solution $x = (x_1, x_2) = (2, -1)$. The same geometric visualization holds if A is any 2×2 matrix where the first column is nonzero and the second column is not a scalar multiple of the first. In

this situation, we can linearly combine the two columns to uniquely obtain any desired $b \in \mathbb{R}^2$. This corresponds to the fact that if the two columns of A are linearly independent, then A is invertible and $Col(A) = \mathbb{R}^2$ (Theorem 5.4).

D.1.2. Row perspective

View A and b according to their rows

$$A = \begin{bmatrix} - & r_1^\top & - \\ - & r_2^\top & - \\ & \vdots & \\ - & r_m^\top & - \end{bmatrix}, \quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_m \end{bmatrix}$$

where $r_i \in \mathbb{R}^n$ for all $i \in \{1, \ldots, n\}$. Solving $Ax = b$ is equivalent to finding $x \in \mathbb{R}^n$ to simultaneously satisfy all of the following m equations:

$$r_1^\top x = b_1 \tag{D.2}$$

$$r_2^\top x = b_2 \tag{D.3}$$

$$\vdots$$

$$r_m^\top x = b_m \tag{D.4}$$

Each equation $i \in \{1, \ldots, m\}$ has the form $r_i^\top x = b_i$. If $r_i = 0$ (the all-zero vector) then the equation reduces to $0 = b_i$ and is true if and only if $b_i = 0$. If $r_i \neq 0$ then the equation $r_i^\top x = b_i$ always has at least one solution (regardless of $b_i \in \mathbb{R}$), and in fact the set of all $x \in \mathbb{R}^n$ that satisfy $r_i^\top x = b_i$ forms a $(n-1)$-dimensional *hyperplane* within $\mathbb{R}^n$. (Hyperplanes are formally defined in the next section.) Then, a vector $x \in \mathbb{R}^n$ satisfies $Ax = b$ if and only if x is in the intersection of the m hyperplanes defined in the system of equations (D.2)-(D.4).

Consider the same example in the previous subsection:

$$A = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad b = \begin{bmatrix} 1 \\ 3 \end{bmatrix}$$

The system of equations for $Ax = b$ is

$$x_1 + x_2 = 1 \tag{D.5}$$

$$x_1 - x_2 = 3 \tag{D.6}$$

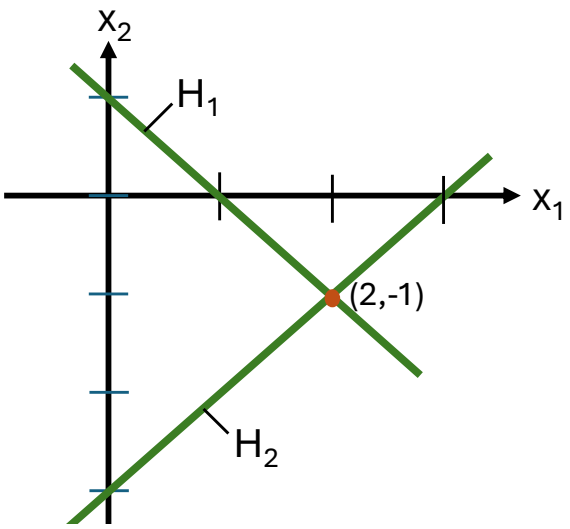


FIGURE D.2. Row perspective: The two intersecting lines (hyperplanes) H_1 and H_2 for the system of equations (D.5)-(D.6). The axes (x_1, x_2) are different from the two axes for the column perspective Fig. D.1.

The reader can draw the two hyperplanes in the 2-dimensional plane $\mathbb{R}^2$:

$$H_1 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 + x_2 = 1\}$$
$$H_2 = \{(x_1, x_2) \in \mathbb{R}^2 : x_1 - x_2 = 3\}$$

From Fig. D.2, one can see that H_1 and H_2 are distinct nonparallel lines in $\mathbb{R}^2$. They intersect at exactly one point, which is unique solution $x = (x_1, x_2) = (2, -1)$.

Suppose this example is changed so that A is some other 2×2 matrix with real entries, and b is some other vector in $\mathbb{R}^2$. For simplicity, assume neither row of A is the all-zero row. The corresponding system of equations produces two hyperplanes H_1 and H_2, both being lines in $\mathbb{R}^2$. There are three possibilities:

- The two lines H_1 and H_2 are distinct and nonparallel: Then there is one and only one point in their intersection. This gives one unique solution to $Ax = b$. This occurs when A is an invertible matrix.
- The two lines H_1 and H_2 are distinct and parallel: Then there is no solution to $Ax = b$. This occurs when A is not an invertible matrix (so row 2 is a scalar multiple of row 1) and $b \notin Col(A)$.
- The two lines H_1 and H_2 are in fact the same line (so $H_1 = H_2$): Then there are infinitely many solutions to $Ax = b$, each point of H_1 being a solution. This occurs when A is not an invertible matrix (so row 2 is a scalar multiple of row 1) and $b \in Col(A)$.

D.2. Hyperplanes

Let n be a positive integer. Fix $\mu \in \mathbb{R}^n$ such that $\mu \neq 0$. Fix $c \in \mathbb{R}$. A *hyperplane in* $\mathbb{R}^n$ is a set $H \subseteq \mathbb{R}^n$ of the form

$$H = \{x \in \mathbb{R}^n : \mu^\top x = c\}$$

Thus, every hyperplane in $\mathbb{R}^n$ is defined by a pair (μ, c) where $\mu \in \mathbb{R}^n$, $\mu \neq 0$, $c \in \mathbb{R}$. Note that μ has the form

$$\mu = (\mu_1, \ldots, \mu_n)$$

where there is at least one $i \in \{1, \ldots, n\}$ such that $\mu_i \neq 0$.

A hyperplane $H \subseteq \mathbb{R}^n$ divides $\mathbb{R}^n$ into two closed *half-spaces* U and L. Specifically, letting (μ, c) be the parameters for H, we have

$$U = \{x \in \mathbb{R}^n : \mu^\top x \geq c\}$$
$$L = \{x \in \mathbb{R}^n : \mu^\top x \leq c\}$$

The hyperplane H is the intersection of the half-spaces U and L, and is on the boundary of both of these half-spaces. The sets U and L are often called the *upper half-space* and *lower half-space*. The terms *upper* and *lower* are used only as labels. A hyperplane H that is characterized by parameters (μ, c) can equally be characterized by $(-\mu, -c)$, and the roles of U and L are swapped when defining *upper* and *lower* from (μ, c) versus $(-\mu, -c)$.

D.2.1. Applications of hyperplanes

Hyperplanes are used extensively in the field of convex optimization (see, for example, Bertsekas, Nedic, and Ozdaglar (2003), Boyd and Vandenberghe (2004)). They are used to place lower bounds on convex functions $f : \mathbb{R}^n \to \mathbb{R}$. For example, it can be shown that if $f : \mathbb{R}^n \to \mathbb{R}$ is a convex function, then for each $a \in \mathbb{R}^n$, there is a vector $f'(a) \in \mathbb{R}^n$, called a *subgradient of f at point a*, such that:

$$f(x) \geq f(a) + f'(a)^\top (x - a) \quad \forall x \in \mathbb{R}^n$$

So for each $x \in \mathbb{R}^n$, the point $(f(x), x) \in \mathbb{R}^{n+1}$ lies in the upper half-space of the hyperplane $H \subseteq \mathbb{R}^{n+1}$ defined by (μ, c), where $\mu = (1, -f'(a))$ and $c = f(a) - f'(a)^\top a$. That is

$$\begin{aligned} H &= \{(g, x) : g \in \mathbb{R}, x \in \mathbb{R}^n, \mu^\top (g; x) = c\} \\ &= \{(g, x) : g \in \mathbb{R}, x \in \mathbb{R}^n, g = f(a) + f'(a)^\top (x - a)\} \end{aligned}$$

Also, a well known *hyperplane separation theorem* in the field of convex optimization says that if X is a convex subset of $\mathbb{R}^n$ and if $y \in \mathbb{R}^n$ is a point such that either $y \notin X$, or y

is on the boundary of X, then there exists a hyperplane in $\mathbb{R}^n$ that passes through y and contains the set X on one side. That is, there is a vector $\mu \in \mathbb{R}^n$, $\mu \neq 0$ such that

$$\mu^\top x \geq \mu^\top y \quad \forall x \in X$$

This result is crucial to the theory of *Lagrange multipliers* and the theory of *duality*.

In linear algebra, the set of all $x \in \mathbb{R}^n$ that satsify the equation $Ax = b$ is equal to the intersection of hyperplanes, as described by the system of equations (D.2)-(D.4).

D.2.2. Equivalent definition of hyperplane

LEMMA D.1. *Fix $n \in \mathbb{N}$. A set $H \subseteq \mathbb{R}^n$ is a hyperplane if and only if it is a $(n-1)$-dimensional affine subset of $\mathbb{R}^n$, that is*

$$H = S + w = \{s + w : s \in S\}$$

for some $(n-1)$-dimensional subspace $S \subseteq \mathbb{R}^n$ and some vector $w \in \mathbb{R}^n$.

PROOF. See next subsection. □

Thus:

1. A hyperplane in $\mathbb{R}$ is a single point of dimension 0. That point is the boundary that separates two half-lines in $\mathbb{R}$.
2. A hyperplane in $\mathbb{R}^2$ is a 1-dimensional line. That line is the boundary that separates two half-planes in $\mathbb{R}^2$. Imagine the line drawn on an infinite sheet of paper. It separates the paper into two half-papers: everything on one side of the line, and everything on the other. You cannot pass from one side of the paper to the other without hitting the line.
3. A hyperplane in $\mathbb{R}^3$ is a 2-dimensional plane. That plane is the boundary that separates two half-spaces in $\mathbb{R}^3$. Imagine the plane as an infinite sheet of paper. It separates the known 3-dimensional universe into two half-universes: everything on one side of the paper, and everything on the other. You cannot pass from one side of the universe to the other without hitting the paper. Even then, Gandalf the wizard may be waiting for you on the paper (he is hard to pass).
4. A hyperplane in $\mathbb{R}^4$ is a 3-dimensional space. That space is the boundary that separates two half-spaces of $\mathbb{R}^4$. You cannot pass from one half-space of $\mathbb{R}^4$ to the other without hitting the 3-dimensional space that forms our hyperplane.

D.2.3. Proof of Lemma D.1

Fix $n \in \mathbb{N}$. For the forward direction of Lemma D.1, let H be a hyperplane in $\mathbb{R}^n$ defined by (μ, c), where $\mu \in \mathbb{R}^n$, $\mu \neq 0$, $c \in \mathbb{R}$. We want to show $H = S + w$ for some $(n-1)$-dimensional

subspace S and some $w \in \mathbb{R}^n$. Define the $1 \times n$ matrix A as follows:

$$A = \begin{bmatrix} \mu_1 & \mu_2 & \cdots & \mu_n \end{bmatrix}$$

Then

$$\begin{aligned} H &= \{x \in \mathbb{R}^n : \mu^\top x = c\} \\ &= \{x \in \mathbb{R}^n : Ax = c\} \\ &= \underbrace{x^*}_{w} + \underbrace{Null(A)}_{S} \end{aligned}$$

where x^* is any particular solution to $Ax = b$ (recall Theorem 4.1). It remains to show: (i) A particular solution x^* always exists; (ii) $Null(A)$ is $(n-1)$-dimensional. Both can be shown using Gaussian elimination to find all solutions to $Ax = c$: Form the augmented matrix:

$$\left[\begin{array}{cccc|c} \mu_1 & \mu_2 & \cdots & \mu_n & c \end{array}\right]$$

This augmented matrix has just one row, and not all μ_i values are zero, *so it is already in echelon form*. There is one pivot that lies in a column $i \in \{1, \ldots, n\}$, where i is the smallest index for which $\mu_i \neq 0$. It follows that $Ax = b$ can always be solved (regardless of $c \in \mathbb{R}$). Also, since there is 1 pivot, there are $n-1$ free variables, so $\dim Null(A) = n - 1$.

Another way to see that $Null(A)$ has dimension $n-1$ is to observe that A has size $1 \times n$ and so, by Lemma 5.4, we have

$$\text{rank}(A) \leq \min\{1, n\} = 1$$

Therefore, $\text{rank}(A) \in \{0, 1\}$. But A is not an all-zero matrix so $\text{rank}(A) \neq 0$. It follows that $\text{rank}(A) = 1$, so the rank-nullity theorem ensures $\dim Null(A) = n - 1$.

For the reverse direction in Lemma D.1: Suppose $H = S + w$ for some $(n-1)$-dimensional subspace $S \subseteq \mathbb{R}^n$ and some $w \in \mathbb{R}^n$. We want to show $H = \{x : \mu^\top x = c\}$ for some $\mu \in \mathbb{R}^n$, $\mu \neq 0$ and some $c \in \mathbb{R}$. Let $\{v_1, \ldots, v_{n-1}\}$ be vectors in $\mathbb{R}^n$ that form a basis for S. We first find a vector $\mu \in \mathbb{R}^n$, $\mu \neq 0$ such that $v_i^\top \mu = 0$ for all $i \in \{1, \ldots, n-1\}$. This is equivalent to finding a nontrivial solution to the matrix equation $V\mu = 0$, where V is the $(n-1) \times n$ matrix formed by the rows $v_i^\top$:

$$V = \begin{bmatrix} — & v_1^\top & — \\ — & v_2^\top & — \\ & \vdots & \\ — & v_{n-1}^\top & — \end{bmatrix}$$

Since V is a wide matrix, it has a nontrivial null space. Let μ be any nonzero vector in $Null(V)$. So $\mu \in \mathbb{R}^n$, $\mu \neq 0$, and $v_i^\top \mu = 0$ for all $i \in \{1, ..., n-1\}$. Define

$$c = \mu^\top w \tag{D.7}$$

$$\tilde{H} = \{x : \mu^\top x = c\} \tag{D.8}$$

It suffices to show $H = \tilde{H}$.

First suppose $x \in H$. So $x = s + w$ for some $s \in S$. We want to show $x \in \tilde{H}$. Since S has basis $\{v_1, ..., v_{n-1}\}$, s is a linear combination of these basis vectors and

$$x = \sum_{i=1}^{n-1} \theta_i v_i + w$$

for some scalars θ_i. Multiplying the above equation on the left by $\mu^\top$ gives

$$\begin{aligned}\mu^\top x &= \sum_{i=1}^{n-1} \theta_i \underbrace{\mu^\top v_i}_{0} + \mu^\top w \\ &\overset{(a)}{=} \mu^\top w \\ &\overset{(b)}{=} c\end{aligned}$$

where (a) uses $\mu^\top v_i = v_i^\top \mu = 0$ for all $i \in \{1, ..., n-1\}$; (b) uses (D.7). Thus, $\mu^\top x = c$ so the definition of $\tilde{H}$ in (D.8) implies $x \in \tilde{H}$.

Now suppose $x \in \tilde{H}$. We want to show $x \in H$. That is, we want to show $x = s + w$ for some $s \in S$. Since $x \in \tilde{H}$ we know from (D.7)-(D.8):

$$\mu^\top x = c = \mu^\top w$$

That is

$$\mu^\top (x - w) = 0$$

Define A as the $1 \times n$ vector $\mu^\top$. Then

$$A(x - w) = 0 \implies x - w \in Null(A)$$

We know that $\text{rank}(A) = 1$ and $\dim Null(A) = n - 1$. On the other hand, the $n - 1$ linearly independent vectors $\{v_1, ..., v_{n-1}\}$ are also in $Null(A)$ because $Av_i = \mu^\top v_i = 0$ for all $i \in \{1, ..., n-1\}$. So $\{v_1, ..., v_{n-1}\}$ are $(n-1)$ linearly independent vectors in the $(n-1)$-dimensional subspace $Null(A)$, so they must be a basis for $Null(A)$. It follows that $x - w$ is

a linear combination of $\{v_1, ..., v_{n-1}\}$, that is

$$x - w = \sum_{i=1}^{n-1} \theta_i v_i$$

for some scalars θ_i. Define $s = \sum_{i=1}^{n-1} \theta_i v_i$. Since $\{v_1, ..., v_{n-1}\}$ is a basis for S, it follows that $s \in S$ and $x = w + s$.

D.3. Norm

For $x, y \in \mathbb{R}^n$, define the *inner product*

$$\langle x, y \rangle = x^\top y = \sum_{i=1}^{n} x_i y_i$$

This is the *standard inner product* for $\mathbb{R}^n$ (other inner products are considered in Chapter 8). This standard inner product is identical to the dot product between vectors x and y in $\mathbb{R}^n$. For $x \in \mathbb{R}^n$ define the *Euclidean norm* of x, also called the L_2 *norm*, by

$$||x|| = \sqrt{x^\top x} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

The norm $||x||$ is the *length* of vector x, also called the *magnitude* of x, being the Euclidean distance between the vector 0 and vector x. That is, $||x||$ is the distance that x is away from the origin of $\mathbb{R}^n$.

D.3.1. Relation to Pythagorean theorem

It is often easier to work with the squared norm:

$$||x||^2 = x^\top x = \sum_{i=1}^{n} x_i^2 \tag{D.9}$$

This formula for the squared norm is consistent with the Pythagorean theorem (see Fig. D.3a). Specifically, for $x = (x_1, x_2) \in \mathbb{R}^2$ we can plot (x_1, x_2) on the 2-dimensional plane and draw the right triangle. The square of the hypotenuse is equal to the sum of squares of the remaining two sides of the right triangle, so

$$||x||^2 = x_1^2 + x_2^2 \qquad \text{[the formula (D.9) for } n = 2\text{]}$$

For three dimensions, consider $x = (x_1, x_2, x_3) \in \mathbb{R}^3$. Define $y = (x_1, x_2, 0)$. Then y lies on the 2-dimensional plane defined by the first and second axes for $\mathbb{R}^3$ and the Pythagorean

theorem gives (see Fig. D.3b):

$$||y||^2 = x_1^2 + x_2^2$$

On the other hand, vector y forms a right triangle together with the origin and vector x (with the line segment from 0 to x being the hypotenuse), so the Pythagorean theorem gives

$$||x||^2 = ||y||^2 + x_3^2 = x_1^2 + x_2^2 + x_3^2 \qquad \text{[the formula (D.9) for } n = 3\text{]}$$

In the same way, the formula (D.9) for $n > 3$ can be viewed as expressing the length of the vector $x \in \mathbb{R}^n$ by repeatedly using the 2-dimensional Pythagorean theorem.

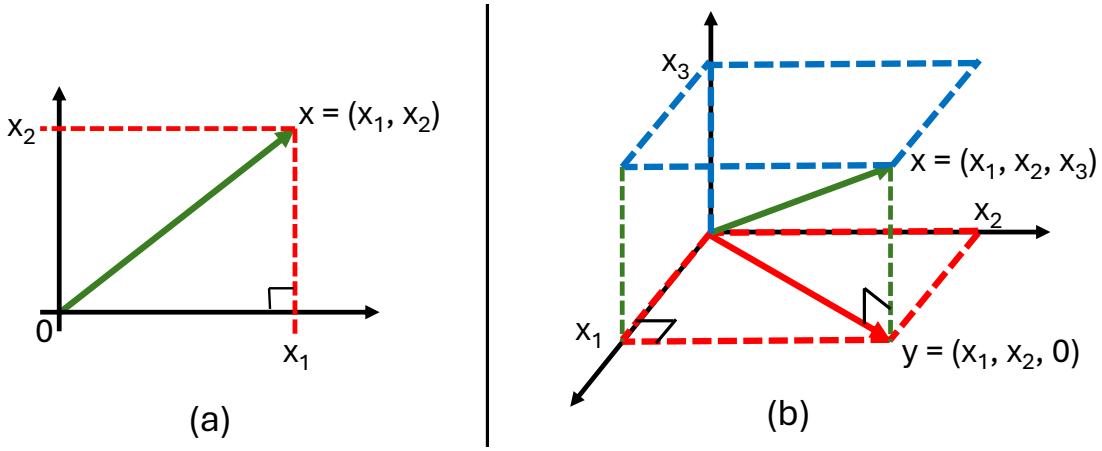


FIGURE D.3. An illustration of the Pythagorean theorem in relation to $||x||^2$ for 2 and 3 dimensions. For (a), $||x||^2 = x_1^2 + x_2^2$. For (b), $||y||^2 = x_1^2 + x_2^2$ and $||x||^2 = ||y||^2 + x_3^2$.

D.3.2. Norm-square of sum

For $x, y \in \mathbb{R}^n$ we have

(D.10) $$||x + y||^2 = ||x||^2 + ||y||^2 + 2x^\top y$$

To prove (D.10), fix $x, y \in \mathbb{R}^n$. Then

$$\begin{aligned} ||x + y||^2 &= (x + y)^\top (x + y) \\ &= x^\top x + x^\top y + y^\top x + y^\top y \\ &= ||x||^2 + ||y||^2 + 2x^\top y \quad \text{[since } x^\top y = y^\top x\text{]} \end{aligned}$$

Suppose x and y are nonzero vectors in $\mathbb{R}^n$. Use x and y to form a triangle with side lengths $||x||$, $||y||$, $||x + y||$ (see Fig. D.4a). This is a right triangle if and only if x and y are

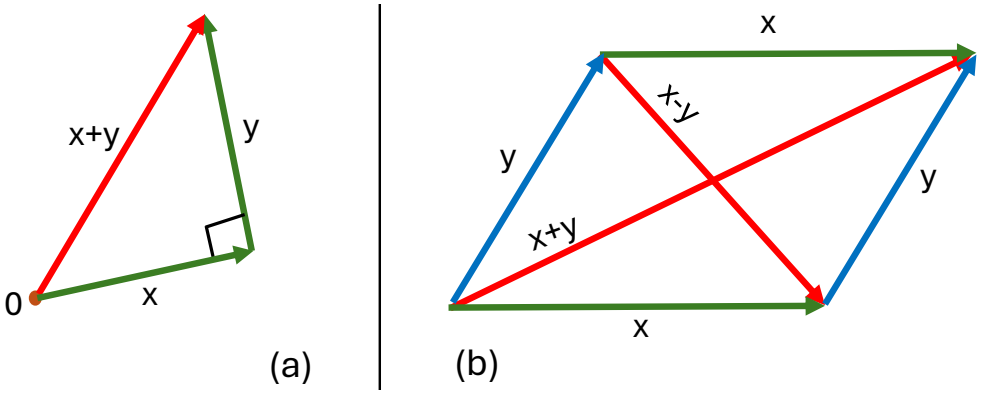


FIGURE D.4. (a) An illustration of $||x + y||$ when x and y are orthogonal; (b) An illustration of the parallelogram law. The side lengths of the parallelogram are $||x||$ and $||y||$. The lengths of the diagonals of the parallelogram are $||x + y||$ and $||x - y||$.

perpendicular. In view of Fig. D.4a, by the Pythagorean theorem we have

$$\begin{aligned}(x \text{ and } y \text{ are perpendicular}) &\iff ||x||^2 + ||y||^2 = ||x + y||^2 \\ &\iff x^\top y = 0\end{aligned}$$

where the final line follows by (D.10). Thus, the inner product $x^\top y$ is geometrically useful because two nonzero vectors $x, y \in \mathbb{R}^n$ are perpendicular if and only if $x^\top y = 0$. The term *orthogonal* is a near-synonym for *perpendicular,* but allows one or both vectors to be zero.

DEFINITION D.1. *Two vectors x and y in $\mathbb{R}^n$ are said to be* orthogonal *(with respect to the standard inner product for $\mathbb{R}^n$) if $x^\top y = 0$.*

D.3.3. Parallelogram law

For $x, y \in \mathbb{R}^n$ we have the *parallelogram law*:

$$||x + y||^2 + ||x - y||^2 = 2\,||x||^2 + 2\,||y||^2$$

The proof uses the norm-square of sum formula:

$$\begin{aligned}||x + y||^2 + ||x - y||^2 &= \left(||x||^2 + ||y||^2 + 2x^\top y\right) + \left(||x||^2 + ||-y||^2 + 2x^\top(-y)\right) \\ &= 2\,||x||^2 + 2\,||y||^2\end{aligned}$$

Geometrically, the two vectors x, y can be used to define the sides of a parallelogram with side lengths $||x||$ and $||y||$ and diagonals of lengths $||x + y||$ and $||x - y||$ (see Fig. D.4b). With this interpretation, the parallelogram law says the sum of the squared lengths of the diagonals of the parallelogram is equal to twice the sum of the squared side lengths.

D.3.4. Angle

The angle θ between two nonzero vectors $x, y \in \mathbb{R}^n$ obeys the formula:

$$x^\top y = ||x|| \cdot ||y|| \cos(\theta) \tag{D.11}$$

If θ is a right angle (so either $\theta = \pi/2$ or $\theta = -\pi/2$) then x and y are orthogonal.

The cosine formula for angle between two nonzero vectors $x, y \in \mathbb{R}^n$ can be understood by viewing the three points $0, x, y$ as vertices of a triangle (see Fig. D.5). The angle θ is the angle between the line segment that points from 0 to x, and the line segment that points from 0 to y. To find $\cos(\theta)$, we break y into perpendicular components: $y = ax + w$ where $a \in \mathbb{R}$ and w is orthogonal to ax. For simplicity, assume we have an acute angle, so $-\pi/2 < \theta < \pi/2$ and $a > 0$. This means the line segment from 0 to y is the hypotenuse of a right triangle with side lengths $||ax||$ and $||w||$ (see Fig. D.5b), where the line segment of size $||ax||$ is adjacent to the angle θ, so

$$\cos(\theta) = \frac{||ax||}{||y||} = \frac{a\,||x||}{||y||} \tag{D.12}$$

The final equality of (D.12) holds because we assume $a > 0$, so $||ax|| = a\,||x||$ (if we had $a < 0$ then we would get $||ax|| = -a\,||x||$).

In the language of projection theory (Chapter 9), we can find $a \in \mathbb{R}$ and $w \in \mathbb{R}^n$ by *projecting y onto $Span\{x\}$*. For this, we find the unique $a \in \mathbb{R}$ for which x is orthogonal to $y - ax$:

$$(x^\top (y - ax) = 0) \iff (x^\top y - ax^\top x = 0) \iff (x^\top y = a\,||x||^2)$$

So $a = (x^\top y)/\,||x||^2$. Substituting this expression for a into (D.12) yields formula (D.11).

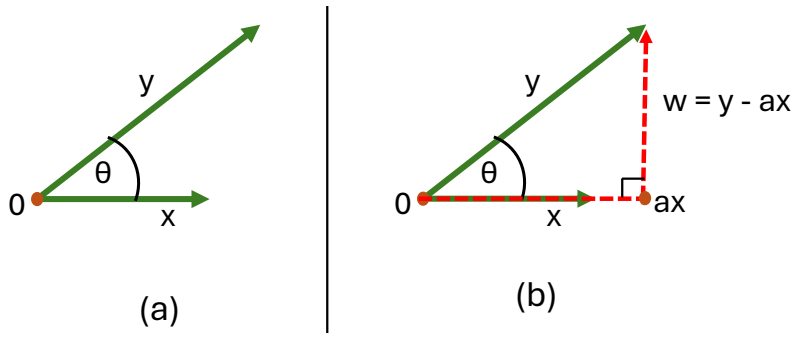


FIGURE D.5. (a) An illustration of an (acute) angle θ between vectors x and y; (b) Breaking y into perpendicular components for $\cos(\theta) = ||ax|| \,/\, ||y||$.

D.4. Rotation matrix for $\mathbb{R}^2$

Fix $\theta \in \mathbb{R}$. Define the *rotation matrix*

$$A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{D.13}$$

Consider a function $T : \mathbb{R}^2 \to \mathbb{R}^2$ defined by

$$T(x_1, x_2) = A \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \quad \forall (x_1, x_2) \in \mathbb{R}^2$$

The function T maps an *input vector* $x \in \mathbb{R}^2$ to an *output vector* $y \in \mathbb{R}^2$, where $y = Ax$. The vector y is a rotated version of x, rotated clockwise about the origin by the angle θ. For example (see Fig. D.6):

$$A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}$$
$$A \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix}$$

It holds that T maps any point $x = (x_1, x_2)$ in the unit circle $\{(x_1, x_2) \in \mathbb{R}^2 : x_1^2 + x_2^2 = 1\}$ to a vector y on the same unit circle, only rotated by angle θ.

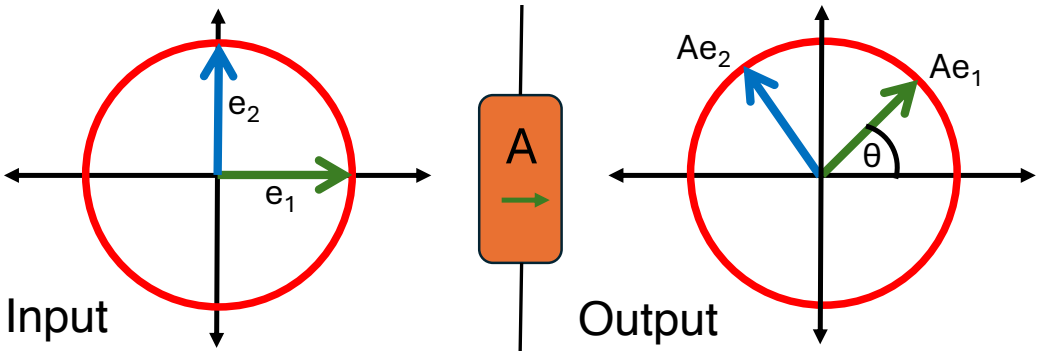


FIGURE D.6. An illustration of left-multiplication by the rotation matrix A in (D.13). The point $e_1 = (1, 0)$ maps to $Ae_1 = (\cos(\theta), \sin(\theta))$. The point $e_2 = (0, 1)$ maps to $Ae_2 = (-\sin(\theta), \cos(\theta))$. Points on the unit circle map to rotated points on the unit circle.

D.4.1. Left-multiplication by rotation matrix A preserves distances

Let x, w be two points in the input space $\mathbb{R}^2$. Assume $x \neq w$ and let $||x - w||$ denote the distance between these points. Let y, z be the corresponding points in the output space $\mathbb{R}^2$ after left-multiplication by the matrix A in (D.13) (see Fig. D.7):

$$y = Ax \quad , \quad z = Aw$$

We want to show that distances are preserved, so that $||y - z|| = ||x - w||$. We have

$$\begin{aligned} ||y - z||^2 &= (y - z)^\top (y - z) \\ &= (Ax - Aw)^\top (Ax - Aw) \\ &= (A(x - w))^\top A(x - w) \\ &= (x - w)^\top A^\top A(x - w) \end{aligned}$$

Now observe that

$$A^\top A = \underbrace{\begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}}_{A^\top} \underbrace{\begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}}_{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I$$

which uses the fact $\cos(\theta)^2 + \sin(\theta)^2 = 1$ for all $\theta \in \mathbb{R}$. Substituting this into the previous equality gives

$$||y - z||^2 = (x - w)^\top A^\top A(x - w) = (x - w)^\top I(x - w) = (x - w)^\top (x - w) = ||x - w||^2$$

Thus, left-multiplication by the A matrix preserves distances.

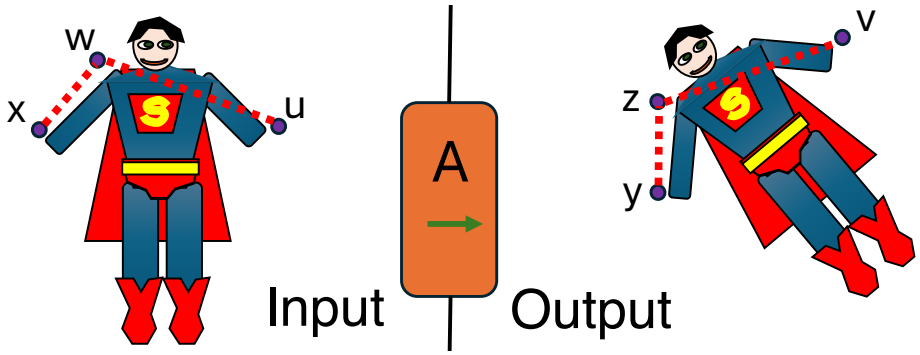


FIGURE D.7. An illustration that left-multiplication by a rotation matrix with angle θ preserves the distance between input points x and y and output points $y = Ax$ and $z = Aw$. Also, the angle associated with points x, w, u is the same as the angle for y, z, v.

D.4.2. Left-multiplication by rotation matrix A preserves angles

Let x, w, u be three vectors in the input space $\mathbb{R}^2$ that form vertices of a triangle. Let y, z, v be the corresponding vectors in the output space $\mathbb{R}^2$ (see Fig. D.7).

$$y = Ax, \quad z = Aw, \quad v = Au$$

The angle θ between vectors $(x - w)$ and $(u - w)$ can be determined using the cosine formula for inner products.

$$(x - w)^\top (u - w) = ||x - w|| \cdot ||u - w|| \cos(\theta)$$

Likewise, the angle β between $(y - z)$ and $(v - z)$ satisfies

$$(y - z)^\top (v - z) = ||y - z|| \cdot ||v - z|| \cos(\beta)$$

We already know $||y - z|| = ||x - w||$ and $||v - z|| = ||u - w||$. To show angles are preserved, it suffices to show inner products are preserved:

$$\begin{aligned}
(y - z)^\top (v - z) &= (Ax - Aw)^\top (Au - Aw) \\
&= (A(x - w))^\top A(u - w) \\
&= (x - w)A^\top A(u - w) \\
&= (x - w)I(u - w) \\
&= (x - w)^\top (u - w)
\end{aligned}$$

An alternative (and simpler) proof that the matrix A preserves angles is to note that we already know the distances between any pair of the three points x, w, u in the input space are equal to the corresponding distances between that pair of points from y, z, v in the output space. The Side-Side-Side postulate of geometry states that two triangles that have exactly the same three side lengths must be congruent. Neely would like to thank a student in the Fall 2026 semester of linear algebra (EE 510 at USC) for this simpler proof.

D.5. Unitary matrices and orthogonality

Vectors $\{x_1, \ldots, x_n\}$ in $\mathbb{R}^m$ are said to be *orthogonal* if

$$x_i^\top x_j = 0 \quad \forall i \neq j$$

Vectors $\{x_1, \ldots, x_n\}$ in $\mathbb{R}^m$ are said to be *orthonormal* if they are orthogonal and

$$||x_i|| = 1 \quad \forall i \in \{1, \ldots, n\}$$

Let A be a real $m \times n$ matrix with orthonormal columns. Then

$$A^\top A = I \quad \text{[the } n \times n \text{ identity matrix]}$$

This holds because the ith diagonal entry of matrix $A^\top A$ is the inner product between column i of A and itself, which is 1 because that column vector is normal. On the other hand, the (i, j) entry of $A^\top A$, for $i \neq j$, is the inner product between columns i and j of A, which is zero because the columns are orthogonal. Examples of nonsquare real matrices A that satisfy $A^\top A = I$ are in Chapter 2.10.2 (see also Appendix B).

The fact $A^\top A = I$ can be used to prove that the columns of A are linearly independent. Thus, if vectors are orthonormal, they must be linearly independent. Therefore, it is impossible to find a collection of more than n orthonormal vectors in $\mathbb{R}^n$. All real matrices A that have orthonormal columns must be either square or tall.

Left-multiplication by a matrix with orthonormal columns preserves lengths: Let A be a real $m \times n$ matrix with orthonormal columns. Fix $x \in \mathbb{R}^n$ and let $y = Ax$, so $y \in \mathbb{R}^m$. Then

$$||y||^2 = y^\top y = (Ax)^\top Ax = x^\top A^\top Ax = x^\top Ix = x^\top x = ||x||^2$$

A (real) *unitary matrix* is a square matrix with real entries and orthonormal columns. The 2×2 rotation matrix A in (D.13) is a unitary matrix. If A is a real unitary matrix then we must have $A^\top A = I$, and since A is square it means $A^{-1} = A^\top$. Hence, we also get $AA^\top = I$. In other words, the inverse of a real unitary matrix is equal to its transpose, and this transpose is itself a unitary matrix. Therefore, a real and square matrix has orthonormal columns if and only if it has orthonormal rows.

D.5.1. Terminology

A real unitary matrix is also called an *orthogonal matrix*. Some people like to make the following terminological distinction:

- An *orthogonal matrix* is a real square matrix A such that $A^\top A = I$.
- A *unitary matrix* is a square matrix A with real or complex entries such that $A^H A = I$, where A^H denotes the conjugate transpose of A (matrix A^H is formed by taking the complex conjugates of all entries of $A^\top$). If A has all real entries then $A^H = A^\top$.

With this distinction, an orthogonal matrix is always unitary, but a unitary matrix is more general because it allows entries in the field $\mathbb{C}$ of complex numbers. The current chapter

deals exclusively with the real field, so unitary matrices and orthogonal matrices are one and the same. There appears to be no name for a (possibly nonsquare) matrix with orthonormal columns, so we simply call it a matrix with orthonormal columns. Such matrices must be either square or tall. When working in the real field, two vectors $x, y \in \mathbb{R}^n$ are defined to be orthogonal if and only if $x^\top y = 0$. When working in the complex field, two vectors $x, y \in \mathbb{C}^n$ are defined to be orthogonal if and only if $x^H y = 0$.

D.6. Realizable rotations in $\mathbb{R}^2$

Do all real 2×2 unitary matrices correspond to rotations in the 2-dimensional plane? Almost. Fix $\theta \in \mathbb{R}$ and consider the following two unitary matrices:

$$A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \quad , \quad B = \begin{bmatrix} -\sin(\theta) & \cos(\theta) \\ \cos(\theta) & \sin(\theta) \end{bmatrix}$$

A is the rotation matrix of (D.13). It corresponds to a rotation by θ that can be physically realized in the 2-dimensional plane. Matrix B is formed by swapping the columns of A. Thus

$$A \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix} \quad , \quad B \begin{bmatrix} 1 \\ 0 \end{bmatrix} = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix}$$
$$A \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} -\sin(\theta) \\ \cos(\theta) \end{bmatrix} \quad , \quad B \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} \cos(\theta) \\ \sin(\theta) \end{bmatrix}$$

Left-multiplication by B first swaps the x and y axes in the 2-dimensional plane, then rotates by θ. It cannot be physically realized in the 2-dimensional plane. The axis swapping has a mirror affect similar to looking at yourself in the mirror: The person in the mirror has a left hand that is the same as your right hand. If you look at a rotated picture on a piece of paper, the corresponding mirrored picture can be viewed as looking from the opposite side of the paper! (Suppose the paper is transparent and the picture can be viewed from both sides.)

It turns out that the determinant of a real unitary matrix is always either 1 or -1 (determinants for general matrices are defined in Chapter 11, and area/volume properties of determinants are discussed in Chapter 11.12). For the 2×2 matrices A and B above we have

$$\det(A) = 1 \quad , \quad \det(B) = -1$$

In the real 2×2 case, a unitary matrix with determinant 1 corresponds to a realizable

rotation by some angle θ, while a unitary matrix with determinant -1 corresponds to a mirrored rotation that cannot be physically realized in the 2-dimensional plane.

It is interesting to note that an unrealizable mirrored rotation in the 2-dimensional plane can be physically realized by opening up a third dimension. Perform the following experiment:

1. Hold your right hand in front of you, so you are looking at the back of your hand. Point your index finger straight up and your thumb to your left. Your finger and thumb form a backwards *L*. Keep your finger and thumb rigid in this experiment.
2. Rotate your hand in the 2-dimensional plane formed by your finger and thumb. You probably will not be able to rotate by an angle more than plus or minus 20 degrees, but that is fine. At each angle of rotation, the thumb remains oriented pointing left with respect to the finger. The same would be true if you could rotate by any angle θ you liked. You cannot make the thumb point right with respect to the finger. Or can you?
3. Start with the position of step 1. Fix our 2-dimensional plane as the one formed by the finger and thumb in this starting position. Now rotate your hand 90 degrees in a direction perpendicular to the 2-dimensional plane, so your index finger points directly away from you (the thumb continues to point directly to your left). Keep going an additional 90 degrees until your finger points down (the thumb continues to point to your left). Your finger and thumb are now back in the original 2-dimensional plane.
4. Rotate your hand 180 degrees in the original 2-dimensional plane. Your are now looking at the palm of your hand. Your index finger is pointing straight up, and *your thumb is pointing to the right*. It makes a true *L*.

This simple experiment is similar to looking at your hand from the opposite side of the 2-dimensional plane.

D.7. Three dimensional rotations

The multiplication of two $n \times n$ real unitary matrices A and B is again a real unitary matrix:

$$(AB)^\top (AB) = B^\top \underbrace{A^\top A}_{I} B = B^\top B = I$$

Similarly, the multiplication of any finite number of $n \times n$ real unitary matrices is again a real unitary matrix. For 3-d rotation in (x, y, z) space we have the following three 3×3 real unitary matrices, each one an adapted version of the 2-dimensional rotation matrix A in (D.13):

- Spin the z-axis by angle θ (leaving all points on the z axis unchanged):

$$A_z(\theta) = \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad , \quad A_z(\theta)^{-1} = A_z(-\theta)$$

- Spin the x-axis by angle α (leaving all points on the x axis unchanged):

$$A_x(\alpha) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos(\alpha) & -\sin(\alpha) \\ 0 & \sin(\alpha) & \cos(\alpha) \end{bmatrix} \quad , \quad A_x(\alpha)^{-1} = A_x(-\alpha)$$

- Spin the y-axis by angle β (leaving all points on the y axis unchanged):

$$A_y(\beta) = \begin{bmatrix} \cos(\beta) & 0 & -\sin(\beta) \\ 0 & 1 & 0 \\ \sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \quad , \quad A_y(\beta)^{-1} = A_y(-\beta)$$

The real unitary matrix associated with first spinning the y-axis by angle β and then spinning the z-axis by angle θ is:

$$\begin{aligned} A_z(\theta)A_y(\beta) &= \begin{bmatrix} \cos(\theta) & -\sin(\theta) & 0 \\ \sin(\theta) & \cos(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos(\beta) & 0 & -\sin(\beta) \\ 0 & 1 & 0 \\ \sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \\ &= \begin{bmatrix} \cos(\theta)\cos(\beta) & -\sin(\theta) & -\cos(\theta)\sin(\beta) \\ \sin(\theta)\cos(\beta) & \cos(\theta) & -\sin(\theta)\sin(\beta) \\ \sin(\beta) & 0 & \cos(\beta) \end{bmatrix} \end{aligned} \tag{D.14}$$

Its inverse matrix is one that performs the negative rotations in the opposite order:

$$(A_z(\theta)A_y(\beta))^{-1} = A_y(-\beta)A_z(-\theta)$$

To specify a rotation, it is not enough to determine how a particular point on the unit sphere $\{(a, b, c) \in \mathbb{R}^3 : a^2 + b^2 + c^2\}$ maps to another desired point on that sphere. This is because there are infinitely many rotations that map that point to its final location. These are formed by starting with one such rotation, then spinning the unit sphere along the line that passes through the origin $(0, 0, 0)$ and that final location point. This spin leaves all points on the line anchored in place, but sweeps all points not on the line through an

infinite number of new locations. A general rotation can be described by picking a line the passes through the origin $(0, 0, 0)$ and spinning about that line according to a certain angle. The spin leaves all points on the line unchanged. It can be shown that concatenating any finite sequence of the basic x-spins, y-spins, z-spins described above results in a rotation that is equivalent to a single spin about an invariant line that passes through the origin.

It is possible to create any rotation by utilizing any two of the above three basic rotations in a sequence of three spins (for example, first spinning z, then spinning x, then spinning z again, see Section D.7.2). We can also create any rotation using any particular ordering of all three types of spins (see Section D.7.2). A real unitary matrix that rotates by angles α, β, θ in the order of spinning x, then spinning y, then spinning z, is

$$A_z(\theta)A_y(\beta)A_x(\alpha) \tag{D.15}$$

with corresponding inverse matrix

$$(A_z(\theta)A_y(\beta)A_x(\alpha))^{-1} = A_x(-\alpha)A_y(-\beta)A_z(-\theta)$$

Notes:

- The rotations are with respect to three absolute (x, y, z) axes that do not change.
- Rotations along the same axis commute, for example

$$A_z(\theta_2)A_z(\theta_1) = A_z(\theta_1 + \theta_2) = A_z(\theta_1)A_z(\theta_2) \quad \forall \theta_1, \theta_2 \in \mathbb{R}$$

- Rotations along different axes do not generally commute, so we typically have

$$A_z(\theta)A_y(\beta) \neq A_y(\beta)A_z(\theta)$$

Thus, the inverse of $A_z(\theta)A_y(\beta)$ is typically *not the same* as replacing (θ, β) with $(-\theta, -\beta)$ in the formula (D.14). This can be understood by seeing where the point $(1, 0, 0)$ lands when left-multiplied by $A_z(\theta)A_y(\beta)$. We first have

$$\begin{bmatrix} \cos(\beta) \\ 0 \\ \sin(\beta) \end{bmatrix} = A_y(\beta) \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

which rotates the point $(1, 0, 0)$ along a circle of radius 1. When we multiply the result by $A_z(\theta)$ (spinning the z axis by angle θ), the point retains its new z-coordinate and rotates its remaining two coordinates in a circle of *radius smaller than or equal to 1*. The smaller radius for this second rotation makes a smaller impact.

- It can be shown that all real 3×3 unitary matrices A have $\det(A) \in \{1, -1\}$ (the determinant formula and its properties are developed for general $n \times n$ matrices in Chapter 11). Matrices that can be formed by starting with the 3×3 identity matrix and then performing a sequence of x-spins, y-spins, and/or z-spins, as described above, are *physically realizable rotation matrices* and have determinant 1. They can be written in the form (D.15). Real unitary matrices that cannot be physically realized in this way have determinant -1. They can be formed by swapping any two columns or any two rows of a physically realizable rotation matrix. They can also be formed by multiplying any row or column of a physically realizable rotation matrix by -1.
- For $\mathbb{R}^3$, a real unitary matrix that corresponds to a physically realizable rotation is a 3×3 real matrix A such that $A^\top A = I$ and $\det(A) = 1$. Such a matrix has orthonormal columns that point in directions consistent with the *right-hand-rule*: If the first and second axes point in directions of our first and second fingers, the third axis points in the direction of the thumb. We call such a matrix a 3-dimensional rotation matrix. A 3-dimensional mirrored rotation corresponds to a real unitary matrix A with $\det(A) = -1$. Similar to the example in D.6, such a mirrored rotation is physically impossible in $\mathbb{R}^3$ but can be accomplished by utilizing a fourth dimension.

D.7.1. Rotations with respect to a different frame of reference

Let $\{e_1, e_2, e_3\}$ be the standard orthonormal basis vectors for $\mathbb{R}^3$, being the three columns of the 3×3 identity matrix:

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} \quad , \quad e_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix} \quad , \quad e_3 = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$$

Vectors $\{e_1, e_2, e_3\}$ provide a *universal reference frame*, also called an *absolute reference frame*, corresponding to fixed x, y, z axes. Each e_i vector is on the 3-dimensional unit sphere. Now suppose we have a different reference frame consisting of some other three orthonormal vectors $\{m_1, m_2, m_3\}$ on the 3-dimensional unit sphere. For example, a character in a video game may be currently oriented in a direction corresponding to reference frame $\{m_1, m_2, m_3\}$. The reference vectors $\{m_1, m_2, m_3\}$ can change over time depending on how the character moves. How can rotations be done relative to reference frame $\{m_1, m_2, m_3\}$? This problem has applications in computer graphics, robotics, and aircraft control.

First assume that reference frame $\{m_1, m_2, m_3\}$ is achievable starting from $\{e_1, e_2, e_3\}$ and performing a physically realizable rotation. Let *reference matrix* M be the physically realizable 3-dimensional rotation matrix (so M is real, $M^\top M = I$, and $\det(M) = 1$). For each

$i \in \{1, 2, 3\}$, assume m_i corresponds to the rotated version of e_i:

$$m_i = Me_i \quad \forall i \in \{1, 2, 3\}$$

This can be written in matrix form as

$$\begin{bmatrix} | & | & | \\ m_1 & m_2 & m_3 \\ | & | & | \end{bmatrix} = M \underbrace{\begin{bmatrix} | & | & | \\ e_1 & e_2 & e_3 \\ | & | & | \end{bmatrix}}_{I} = M$$

In other words, the vectors $\{m_1, m_2, m_3\}$ are the three columns of the rotation matrix M.

We want to obtain a new rotation matrix M_{new} that corresponds to performing rotations in desired directions and by desired angles with respect to the reference frame $\{m_1, m_2, m_3\}$. The simple way to do this is to first pretend the current reference frame is the universal frame $\{e_1, e_2, e_3\}$, find the rotation matrix A that we would want if this really was the case, and then rotate that matrix A by M:

$$\boxed{M_{new} = MA} \tag{D.16}$$

We call the idea of first performing the desired rotation in the universal frame, then left-multiplying by the reference matrix M, the *reference principle*. The equation (D.16) is the corresponding *reference principle equation*. For example,

- Spin the axis m_3 by an angle θ: $M_{new} = MA_z(\theta)$.
- Spin the axis m_1 by an angle α: $M_{new} = MA_x(\alpha)$.
- Spin the axis m_2 by an angle β: $M_{new} = MA_y(\beta)$.

Since matrices that represent spins along different axes do not necessarily commute, care must be taken to obtain a smooth rotation in a direction that mixes the m_1, m_2, m_3 directions. This is easiest to describe by imagining a *flight simulator* scenario. The perspective of the pilot is shown as a 2-dimensional *viewpoint plane* in Fig. D.8. The pilot is looking outward. The reference frame $\{m_1, m_2, m_3\}$ is organized so that m_1 points horizontally to the right, m_2 points vertically up, and m_3 points directly towards the pilot. Imagine the pilot wants to move in a direction that is either up/down, left/right, or diagonal:

- Spin axis m_2 (move right/left) by angle β: $M_{new} = MA_y(\beta)$.
- Spin axis m_1 (move up/down) by angle α: $M_{new} = MA_x(\alpha)$.
- Spin the diagonal $\tilde{m}_2$ (defined by angle θ and moving in $\tilde{m}_1$ direction) by angle β (see

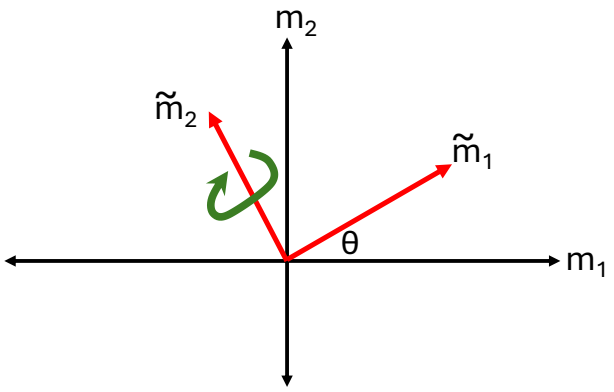


FIGURE D.8. The 2-dimensional viewpoint plane of a pilot looking outward in a flight simulator. The reference frame is given by orthormal vectors $\{m_1, m_2, m_3\}$ in $\mathbb{R}^3$, with corresponding rotation matrix $M = [\ m_1 \quad m_2 \quad m_3\ ]$. To move in the direction $\tilde{m}_1$, we spin axis $\tilde{m}_2$ through a desired angle β, accomplished by rotation matrix $MA_z(\theta)A_y(\beta)$.

Fig. D.8):

$$M_{new} = MA_z(\theta)A_y(\beta)$$

- Spin axis m_3 without changing the 2-d viewpoint plane: $M_{new} = MA_z(\theta)$.

The third bullet uses a matrix $A_z(\theta)A_y(\beta)$ that is itself formed by the reference principle (D.16): Pretend the perspective of the pilot in Fig. D.8 is really the universal reference frame $\{e_1, e_2, e_3\}$, where e_1 points to the right, e_2 points up, e_3 points in toward the pilot. Spinning axis e_3 by angle θ via matrix $A_z(\theta)$ rotates e_1 and e_2 in the 2-d viewpoint plane to align with $\tilde{m}_1$ and $\tilde{m}_2$ (vector e_3 is unchanged). So $A_z(\theta)$ is the matrix of reference frame $\{\tilde{m}_1, \tilde{m}_2, \tilde{m}_3\}$. Since $\tilde{m}_1$ is the direction we want to look, we spin axis $\tilde{m}_2$ by angle β, which is accomplished by pretending $\tilde{m}_2$ is the universal y-axis and using $A_y(\beta)$. We then apply the reference principle (D.16) by left-multiplying by $A_z(\theta)$ to obtain $A_z(\theta)A_y(\beta)$.

The pilot perspective can be mirrored to look from a viewpoint just outside of the aircraft, looking *into* the plane at the pilot, by multiplying the first column of the current M matrix by -1. This is similar to the example of Section D.6: If you look at the back of your hand and form a backward L with finger and thumb, people on the other side see a true L with thumb pointing in the opposite direction. The example of that section was physically realizable in 3-d space because moving your hand according to the given procedure also flips the direction of the third axis (the z-axis). In contrast, in the pilot scenario, multiplying the first column of M by -1 yields a mirrored rotation that is *not* a physically realizable rotation because the z-axis is unchanged.

D.7.2. Achieving any rotation with xyz spins or xyx spins

How can we achieve a rotation matrix that spins about a given line via xyz spins of the type $A_z(\theta)A_y(\beta)A_x(\alpha)$? Fix a normal vector $(a, b, c) \in \mathbb{R}^3$, so $a^2 + b^2 + c^2 = 1$. Consider the line L that passes through $(0, 0, 0)$ and (a, b, c). We want to achieve any rotation that spins line L by a desired angle. First suppose we have some rotation matrix A that maps e_1 to (a, b, c):

$$A \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix} = \begin{bmatrix} a \\ b \\ c \end{bmatrix}$$

Then, the reference principle equation (D.16) ensures all spins of line L can be achieved by first spinning e_1 and then left-multiplying by A, so we use $AA_x(\alpha)$. Specifically:

- Using $A_x(\alpha)$ and sweeping through $\alpha \in [0, 2\pi)$, we spin the x-axis (vector e_1) along all possible angles.
- Using $AA_x(\alpha)$ and sweeping through $\alpha \in [0, 2\pi)$, we spin line L along all possible angles.

Thus, it remains only to construct such a matrix A of the type

$$A = A_z(\theta)A_y(\beta)$$

For this, we just need $A_z(\theta)A_y(\beta)$ to map $(1, 0, 0)$ to (a, b, c). First spin the y-axis, which rotates the vector $(1, 0, 0)$ about a circle of radius 1. This allows the z-coordinate to slide through all possibilities in the interval $[-1, 1]$. Clearly $c \in [-1, 1]$, so we just use an angle β for the y-spin that ensures $(1, 0, 0)$ rotates to a new vector with last coordinate equal to c. Then use a z-spin (which leaves the z-coordinate c unchanged) to find the desired (a, b, c) on the unit sphere.

A similar argument shows how all spins about the same line L can be achieved by $A_x(\gamma)A_y(\beta)A_x(\alpha)$ for some angles α, β, γ. These are so-called xyx spins. In particular, this can be done by finding γ, β such that $A_x(\gamma)A_y(\beta)e_1 = (a, b, c)$.

D.7.3. Replacing and repairing lost reference axes

Suppose we have a vector $a = (a_1, a_2, a_3)$ on the unit sphere $\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = 1\}$. How do we find corresponding vectors $b, c \in \mathbb{R}^3$ such that $\{a, b, c\}$ are orthonormal? Also, how do we ensure the resulting three vectors satisfy the right-hand-rule property, so that creating a matrix M with columns a, b, and c yields a physically realizable rotation matrix?

One way is to first find an orthonormal vector $r \in \mathbb{R}^3$ by finding any nonzero solution to

$$a^\top r = 0$$

We can view $a^\top$ as a 1×3 matrix. This matrix is wide so there are nonzero solutions to $a^\top r = 0$. Find such an $r \neq 0$ (such as by using a very simple form of Gaussian elimination), then normalize it:

$$b = \frac{r}{||r||}$$

Then $\{a, b\}$ are orthonormal vectors in $\mathbb{R}^3$. We need to find a third orthonormal vector: Find any nonzero solution $s \in \mathbb{R}^3$ to

$$As = 0$$

where A is the 2×3 (wide) matrix with rows $a^\top$ and $b^\top$. We can use Gaussian elimination. Then normalize the result

$$c = \frac{s}{||s||}$$

Then $\{a, b, c\}$ are orthonormal in $\mathbb{R}^3$. If they do not satisfy the right-hand-rule (equivalently, if the determinant of the corresponding 3×3 matrix with columns a, b, c is -1), just multiply c by -1 so that $\{a, b, -c\}$ are the desired vectors.

Suppose we have a reference system given by orthonormal vectors $\{m_1, m_2, m_3\}$ in $\mathbb{R}^3$, with corresponding reference matrix M with columns given by m_1, m_2, m_3. The matrix M is a 3×3 matrix with 9 real entries:

$$M = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$$

Suppose we erase the last column:

$$M = \begin{bmatrix} a & b & ?? \\ d & e & ?? \\ g & h & ?? \end{bmatrix}$$

We now claim the missing information can be recovered. This is because the missing

column, call it $x = (x_1, x_2, x_3) \in \mathbb{R}^3$, must be orthogonal to the known first two columns:

$$\underbrace{\begin{bmatrix} a & d & g \\ b & e & h \end{bmatrix}}_{A} \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}}_{x} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

The matrix A has two linearly independent rows (after all, these rows are orthonormal) and so $\text{rank}(A) = 2$ and $\dim Null(A) = 1$ (by the rank-nullity theorem). It follows that

$$Null(A) = Span\{w\}$$

for some $w \in \mathbb{R}^3$ such that $w \neq 0$ and $w \in Null(A)$. Such a vector w can be found by Gaussian elimination. [By the projection theory of Chapter 9, we can also define $w = v - p$ where $v \in \mathbb{R}^3$ is any vector that is not in $Col(A^\top)$ and p is the projection of v onto $Col(A^\top)$.] So the missing column must be a scalar multiple of w. However, we also know the missing column is normal (that is, it has norm 1), so there are only two possibilities:

$$\frac{w}{||w||} \quad , \quad \frac{-w}{||w||}$$

Assuming that the columns of M satisfy the right-hand-rule, we can choose between $w/||w||$ and $-w/||w||$ with full certainty.

If many 3×3 rotation matrices $A_1 A_2 A_3 \cdots A_k$ are multiplied, small truncation errors can add up to make the resulting matrix loose its orthonormality properties. An *approximate* rotation matrix A with columns $\{a, b, c\}$ in $\mathbb{R}^3$ can be repaired to orthonormal columns $\{a_{fresh}, b_{fresh}, c_{fresh}\}$ as follows:

1. Normalize a:
$$a_{fresh} = \frac{a}{||a||}$$
which leaves a unchanged when its norm is already 1.
2. Define
$$b_{fresh} = \frac{b - p}{||b - p||}$$
where p is the projection of b onto $Span\{a_{fresh}\}$ (see Chapter 9):
$$p = (b^\top a_{fresh}) a_{fresh}$$
where $p = 0$ when b is already orthogonal to the unit vector a_{fresh}.

3. Define

$$c_{fresh} = \frac{c - q}{||c - q||}$$

where q is the projection of c onto $Span\{a_{fresh}, b_{fresh}\}$ (see Chapter 9):

$$q = (c^\top a_{fresh})a_{fresh} + (c^\top b_{fresh})b_{fresh}$$

where $q = 0$ if c is already orthogonal to unit vectors a_{fresh} and b_{fresh}.

D.7.4. A brief discussion of quaternions

Consider the situation of the previous subsection. It turns out that all information in the reference matrix M can be efficiently represented by (m_1, θ), where m_1 is the first column of matrix M (so m_1 is a vector in $\mathbb{R}^3$ with all three entries in the interval $[-1, 1]$) and $\theta \in [0, 2\pi)$ is the *angle* associated with the missing orthonormal columns m_2 and m_3 (assuming the columns satisfy the right-hand-rule). Specifically, by choosing any vectors $r, s \in \mathbb{R}^3$ for which $\{m_1, r, s\}$ are orthonormal, and for which the right-hand-rule holds, we can obtain all other possibilities $\tilde{r}, \tilde{s} \in \mathbb{R}^3$ for the second and third columns of M by spinning axis m_1 over the angle θ. The data structure of a *quaternion* is a formal way to represent the information (m_1, θ), with the understanding that this information exactly gives the entire reference matrix M under the right-hand-rule assumption. In comparison to the 3×3 matrix M that stores 9 real numbers, the quaternion stores only four real numbers. A calculus for multiplying quaternions exists to efficiently carry out the reference principle equation (D.16) using the quaternion form of M and A.

D.7.5. Summary

A real unitary matrix is a square matrix A with real entries that satisfies $A^\top A = I$, so the columns of A are orthonormal (the rows are also orthonormal). Now consider such matrices with size 3×3. If $\det(A) = 1$ then the three columns of A satisfy the right-hand-rule and the matrix corresponds to a physically realizable rotation in 3-dimensional space. Else, $\det(A) = -1$ and A represents a mirrored rotation. The reference principle states that we can accomplish a rotation with respect to a given reference frame (characterized by some rotation matrix M) by first performing the desired rotation A with respect to the universal reference frame and then using MA.

APPENDIX E.

Hamel basis

Let V be a vector space over a field $\mathbb{F}$. Let S be a subspace of V. If $\dim(S) = n$ for some $n \in \mathbb{N}$ then a *basis* for S is a list of n vectors $\{v_1, ..., v_n\}$ that are linearly independent and that span S. It is not obvious how to extend the definition of *basis* to the case $\dim(S) = \infty$. Such a basis should be an *infinite* collection of vectors in S. However, we can only form linear combinations by taking a weighted sum of a *finite number* of vectors. That is because infinite sums are not defined for general vector spaces.

This appendix presents the definition of a *Hamel basis*. The definition can be applied to any subspace, including infinite dimensional subspaces. When $\dim(S) < \infty$, the definition of Hamel basis reduces to the standard definition of basis. When $\dim(S) = \infty$, a set theory axiom called the *axiom of choice* ensures that S has a Hamel basis. Therefore (assuming the axiom of choice), *every nontrivial subspace has a Hamel basis*. Further, $\dim(S) = \infty$ if and only if the Hamel basis for S consists of an infinite number of vectors in S.

There are alternative definitions of an infinite basis within mathematics. For example, when working with real-valued functions, some alternative definitions allow basis vectors to participate in *infinite sums*, provided that some convergence criterion is met. This appendix treats only the Hamel basis definition.

E.1. Definition

Let S be a nontrivial subspace of a vector space V (over some field $\mathbb{F}$). Let H be a nonempty set of vectors in S. The set H is said to be a *Hamel basis for S* if:

- Every finite and nonempty subset of vectors in H is linearly independent.
- $Span(H) = S$. That is, every $x \in S$ can be written as a linear combination of vectors in H, so

  (E.1) $$x = \sum_{v \in H_x} a_{v,x} v$$

  where H_x is a nonempty and *finite* subset of H (so $H_x \subseteq H$ and $|H_x| \in \mathbb{N}$), and $a_{v,x} \in \mathbb{F}$ for each $v \in H_x$.

Here we use the usual definition of *linear independent* as applied to a finite set of vectors. If H is a Hamel basis for S, it immediately follows that all vectors in H are nonzero. That is because if $v \in H$ then the subcollection $\{v\}$ must be linearly independent. Since we have defined H as a *set* of vectors (rather than a *list* of vectors), there are no redundantly repeated vectors in H. That is fine because redundantly repeated vectors would not fit the linear independence requirement.

E.1.1. Alternative notation

If H is an uncountably infinite set then its vectors cannot be listed. However, it is standard to use alternative notation x_i to describe a vector in H, where i is an index within some (possibly uncountably infinite) set of indices I. In that case, $x_i \in H$ for all $i \in I$; $H = \cup_{i \in I} \{x_i\}$; the cardinality of I is the same as the cardinality of H. With this notation, the fact that every $x \in S$ is a linear combination of vectors in H means

(E.2) $$x = \sum_{i \in I_x} a_{i,x} v_i$$

for some nonempty and finite set $I_x \subseteq I$ and some scalars $a_{i,x} \in \mathbb{F}$ for all $i \in I_x$. The reader can use either notation (E.1) or (E.2) according to preference.

E.1.2. The case $\dim(S) < \infty$

In the special case $\dim(S) = n$ for some $n \in \mathbb{N}$, any Hamel basis for S must contain exactly n vectors. This is because we cannot have more than n linearly independent vectors in a space of dimension n. Also, we cannot span a subspace of dimension n with fewer than n vectors. Then, the above definition of Hamel basis reduces to the ordinary definition of basis for the finite dimensional subspace S. Also, as in the case of a standard basis, if $S = \{0\}$ then there is no Hamel basis.

E.1.3. Example: Hamel basis for set of polynomials

Let V be the real vector space of polynomial functions $f : \mathbb{R} \to \mathbb{R}$. Define $C = \{1, t, t^2, t^3, ...\}$. In this case, C is a countably infinite list of polynomial functions. We now show that C is a Hamel basis for V.

- Linear independence: Take any finite and nonempty set $B \subseteq C$. Then there is a positive integer n such that $B \subseteq \{1, t, t^2, ..., t^n\}$. We already know $\{1, t, t^2, ..., t^n\}$ is linearly independent, so any finite and nonempty subset of functions in this list is also linearly independent. In particular, the set of functions in B is linearly independent.
- Span: To show that $Span(C) = V$, fix $f \in V$. Then f is a polynomial. Let k be the degree of f. Then $f \in Span\{1, t, ..., t^k\}$.

E.1.4. Example: Hamel basis for infinite sequences

Fix $\mathbb{F}$ as a field. Let $\mathbb{F}^{\mathbb{N}}$ denote the set of infinite sequences of scalars in $\mathbb{F}$:

$$\mathbb{F}^{\mathbb{N}} = \{(x_1, x_2, x_3, ...) : x_i \in \mathbb{F} \quad \forall i \in \mathbb{N}\}$$

Define $S \subseteq \mathbb{F}^{\mathbb{N}}$ as the set of all sequences in $\mathbb{F}^{\mathbb{N}}$ that have at most finitely many nonzero components. That is, $x \in S$ means that $x = (x_1, x_2, x_3, ...)$, $x_i \in \mathbb{F}$ for all $i \in \mathbb{N}$, and there are at most finitely many nonzero x_i values. It can be shown that S is indeed a subspace: The sum of two sequences with at most finitely many nonzeros is again a sequence with at most finitely many nonzeros (also, multiplying a sequence in S by a scalar produces another sequence in S). Define

$$e_1 = (1, 0, 0, 0, ...)$$
$$e_2 = (0, 1, 0, 0, ...)$$
$$e_3 = (0, 0, 1, 0, ...)$$

and so on. Specifically, for each $i \in \mathbb{N}$, $e_i \in S$ is the sequence that is always zero except in entry i (where it is 1). Define the countably infinite set $C = \{e_1, e_2, e_3, ...\}$. We now show that C is a Hamel basis for S:

- Linear independence: Let B be any nonempty and finite subset of C. Then there is a positive integer k such that $B \subseteq \{e_1, e_2, ..., e_k\}$. We already know $\{e_1, e_2, ..., e_k\}$ is linearly independent. Hence, B is also linearly independent.
- Span: To show that $Span(C) = S$, fix $x \in S$. Then $x = (x_1, x_2, ...)$ and there is an index

$k \in \mathbb{N}$ such that $x_i = 0$ whenever $i > k$. It follows that

$$x = \sum_{i=1}^{k} x_i e_i \in Span(C)$$

This shows that $\{e_1, e_2, e_3, ...\}$ is a Hamel basis for the subspace S of all sequences of scalars that have at most finitely many nonzero entries. Therefore, $\{e_1, e_2, e_3, ...\}$ is *not* a Hamel basis for $\mathbb{F}^{\mathbb{N}}$. That is because the always-1 sequence $(1, 1, 1, ...)$ is in $\mathbb{F}^{\mathbb{N}}$ but is not a linear combination of $\{e_1, e_2, e_3, ...\}$. [It does not make sense to say $(1, 1, 1, ...) = \sum_{i=1}^{\infty} e_i$ because infinite sums are not defined for general vector spaces with general fields $\mathbb{F}$. Even if the infinite sum is somehow defined, it would not fit the definition of a linear combination as a *finite sum*.] Exercise 6.30 shows that if C is a collection of vectors in $\mathbb{F}^{\mathbb{N}}$ that span $\mathbb{F}^{\mathbb{N}}$, then C must be uncountably infinite.

E.2. Existence of a Hamel basis

Let V be a vector space over a field $\mathbb{F}$. Let S be a nontrivial subspace. A set theory axiom called the *axiom of choice* can be used to prove that S has a Hamel basis. That is, every nontrivial subspace has a Hamel basis. Furthermore, every Hamel basis of S can be extended to a Hamel basis of V. Specifically, suppose $C \subseteq V$ is a Hamel basis for S. Then there is a set $D \subseteq V$ such that $C \cup D$ is a Hamel basis for V. The proof is beyond the scope of this text.

The axiom of choice ensures existence, but does not provide a way of actually constructing the Hamel basis. In some cases, the special structure of the subspace S allows an explicit Hamel basis to be found (such as for the two examples in the previous section). However, in many cases we can only assert *existence* of a Hamel basis, without any explicit construction. For example, the set $\mathbb{F}^{\mathbb{N}}$ of all infinite sequences of scalars in $\mathbb{F}$ has a Hamel basis, as guaranteed by the axiom of choice, but nobody actually knows how to construct it. Exercise 6.30 shows that any Hamel basis of $\mathbb{F}^{\mathbb{N}}$ must have an uncountably infinite number of basis vectors.

E.3. Uniqueness property

LEMMA E.1. *(Uniqueness property of Hamel basis) Let V be a vector space over a field $\mathbb{F}$. Let H be a Hamel basis for V. Then every $x \in V$ such that $x \neq 0$ can be* uniquely *written*

$$x = \sum_{v \in H_x} a_{v,x} v$$

for some unique finite and nonempty set $H_x \subseteq H$ and nonzero scalars $a_{v,x} \in \mathbb{F}$ for $v \in H_x$.

PROOF. Fix $x \in V$ such that $x \neq 0$. Suppose there are two ways to write x as a linear combination of vectors in H, so

$$x = \sum_{v \in A} a_v v = \sum_{v \in B} b_v v$$

where A and B are two nonempty finite subsets of H; a_v is a nonzero scalar for each $v \in A$; b_v is a nonzero scalar for each $v \in B$. We show that $A = B$ and $a_v = b_v$ for all $v \in A$, so these two ways of expressing x are in fact the same. Define

$$D = A \cup B$$

$$\tilde{a}_v = \begin{cases} a_v & \text{if } v \in A \\ 0 & \text{if } v \notin A \end{cases}$$

$$\tilde{b}_v = \begin{cases} b_v & \text{if } v \in B \\ 0 & \text{if } v \notin B \end{cases}$$

Then D is a finite and nonempty subset of H and

$$0 = \left(\sum_{v \in A} a_v v \right) - \left(\sum_{v \in B} b_v v \right) = \sum_{v \in D} (\tilde{a}_v - \tilde{b}_v) v$$

and since all vectors in the finite set D are linearly independent, we immediately conclude $\tilde{a}_v - \tilde{b}_v = 0$ for all $v \in D$, so

$$\tilde{a}_v = \tilde{b}_v \quad \forall v \in D \tag{E.3}$$

It follows that if $v \in A \cap B$ then

$$a_v = \tilde{a}_v = \tilde{b}_v = b_v$$

Also, we must have $A \cap B = A = B$. Indeed, if there was a vector v such that $v \in A$, $v \notin B$, we would have $\tilde{a}_v = a_v \neq 0$ but $\tilde{b}_v = 0$, contradicting (E.3). A similar contradiction arises if there was a vector $v \in B$ that is not also in A. □

E.4. Existence of inner products for real vector spaces

Let V be a real vector space. If $\dim(V) < \infty$, we already know that V has an inner product. Now suppose $\dim(V) = \infty$. Let H be a Hamel basis for V. Then each $x \in V$ and $y \in V$ such that $x \neq 0$ and $y \neq 0$ can be uniquely written

$$x = \sum_{v \in H_x} a_{v,x} v \quad , \quad y = \sum_{v \in H_y} a_{v,y} v$$

for some finite and nonempty sets $H_x \subseteq H$, $H_y \subseteq H$ and some nonzero scalars $a_{v,x} \in \mathbb{R}$ for $v \in H_x$, $a_{v,y} \in \mathbb{R}$ for $v \in H_y$. Define the function $\langle x, y \rangle$ for all $x, y \in V$ by

$$\langle x, y \rangle = \begin{cases} 0 & \text{if } x = 0 \text{ or } y = 0 \\ \sum_{v \in H_x \cap H_y} a_{v,x} a_{v,y} & \text{if } x \neq 0 \text{ and } y \neq 0 \end{cases}$$

where $H_x, H_y, a_{v,x}, a_{v,y}$ are uniquely defined as above. Note that $\langle x, y \rangle \in \mathbb{R}$ for all $x, y \in V$ because $\sum_{v \in H_x \cap H_y} a_{v,x} a_{v,y}$ is always a finite sum of real numbers. It follows that if u and v are vectors in the Hamel basis H then their unique decompositions are $u = 1 \cdot u$ and $v = 1 \cdot v$ and so

$$\langle u, v \rangle = \begin{cases} 1 & \text{if } u = v \\ 0 & \text{if } u \neq v \end{cases}$$

It can be shown that $\langle x, y \rangle$ is a valid inner product on V. Under this inner product, the vectors in the Hamel basis H can be considered orthonormal.

E.5. Infinite dimensional dual space

Let V be a vector space over field $\mathbb{F}$. Suppose $\dim(V) = \infty$. Let V^* be the dual vector space of all linear transformations $T : V \to \mathbb{F}$. We want to show $\dim(V^*) = \infty$. Fix $n \in \mathbb{N}$. It suffices to construct n linearly independent functions in V^*. (Then $\dim(V^*) \geq n$, and this holds for all $n \in \mathbb{N}$, so $\dim(V^*) = \infty$). Let H be a Hamel basis for V. We know that H contains an infinite number of vectors in V. Each $x \in V$ such that $x \neq 0$ can be uniquely expressed

$$x = \sum_{v \in H_x} a_{v,x} v$$

for some finite and nonempty set $H_x \subseteq H$ and some nonzero scalars $a_{v,x} \in \mathbb{F}$ for $v \in H_x$. For each $u \in H$ define the function $T_u : V \to \mathbb{F}$ by

$$T_u(x) = \begin{cases} 0 & \text{if } u \notin H_x \\ a_{u,x} & \text{if } u \in H_x \end{cases}$$

Note that for $u, v \in H$ the unique expansions are $u = 1 \cdot u$ and $v = 1 \cdot v$ so

$$T_u(v) = \begin{cases} 1 & \text{if } u = v \\ 0 & \text{if } u \neq v \end{cases}$$

It can be shown that T_u is a linear transformation for each $u \in H$. Also, for any finite and nonempty set $B \subseteq H$, the collection $\{T_u : u \in B\}$ can be shown to be linearly independent. So we can find n linearly independent functions in V^* for any $n \in \mathbb{N}$.

It turns out that $\{T_u\}_{u\in H}$ is *not* a Hamel basis for V^*. This is because there are some linear transformations $T : V \to \mathbb{F}$ that cannot be expressed as a linear combination of the transformations in the collection $\{T_u\}_{u\in H}$ (recall that a linear combination is a *finite weighted sum*). While V and V^* both have infinite dimension, the *cardinality* of the Hamel basis for V^* can be shown to be strictly larger than the cardinality of the Hamel basis for V (it is a larger level of infinity). [Recall that if S is a subspace, this text uses the phrase "$\dim(S) = \infty$" to mean only that S cannot be spanned by a finite number of vectors. It means that a Hamel basis for S must have infinitely many vectors, but does not say anything more about the cardinality of the Hamel basis.]

E.6. Things that are not a Hamel basis

Fix $T > 0$. Let V be the real vector space of square integrable functions $f : [-T/2, T/2] \to \mathbb{R}$. Let C be the following countably infinite collection of functions in V (from Chapter 9.12):

$$C = \left\{ \frac{1}{\sqrt{T}}, \quad \left\{ \sqrt{\frac{2}{T}} \cos\left(\frac{2\pi k t}{T}\right) \right\}_{k=1}^{\infty}, \quad \left\{ \sqrt{\frac{2}{T}} \sin\left(\frac{2\pi k t}{T}\right) \right\}_{k=1}^{\infty} \right\}$$

The set C is *not* a Hamel basis for V because $e^t \in V$ (where e^t is treated as having domain $t \in [-T/2, T/2]$), but $e^t \notin Span(C)$. This is because e^t cannot be written as a linear combination (finite weighted sum) of functions in C. Of course, we know from the theory of Fourier series that e^t can be approximated arbitrarily closely by linear combinations of functions in C. In other branches of mathematics, people might say that C is a "complete orthonormal basis for V," but such usage of the term "basis" for an infinite dimensional vector space is not the same as a Hamel basis. [That other definition allows certain types of countably infinite weighted sums of functions in C.]

Similarly, if we define V as the complex vector space of square integrable functions $f : [-T/2, T/2] \to \mathbb{C}$, the countably infinite collection $D = \{e^{2\pi i k t/T}\}_{k=-\infty}^{\infty}$ is *not* a Hamel basis for V. Of course, from Chapter 10 we know that any function $f \in V$ can be approximated arbitrarily closely by linear combinations of functions in D.

Also, a frog is not a Hamel basis. So C, D, and a frog. These are all examples of things that are not a Hamel basis.

Bibliography

Axler, S. 2024. *Linear Algebra Done Right*. 4th ed. Springer.

Bertsekas, D. P., A. Nedic, and A. E. Ozdaglar. 2003. *Convex Analysis and Optimization*. Boston: Athena Scientific.

Boyd, S. and L. Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Lay, David C., Steven R. Lay, and Judi J. McDonald. 2016. *Linear Algebra and its Applications*. 5th ed. Boston: Pearson.

Strang, G. 2023. *Introduction to Linear Algebra, 6th ed.* Wellesly-Cambridge Press.

Acknowledgments

I have significantly benefitted from early versions of the following three linear algebra textbooks:

1. Lay, Lay, and McDonald (2016)
2. Axler (2024)
3. Strang (2023)

Early versions of the first two books above were textbooks for the basic and advanced linear algebra classes I took in college at the University of Maryland (1993-1997). Unfortunately I forget the names of the professors of those courses. The professor who taught from the first book had three distinguishing characteristics: (i) He was a thin bearded man; (ii) He was a good teacher; (iii) He liked to say the number of solutions to $Ax = b$ was "either one, none, or infinitely many." Since then, I have never met anyone with all three characteristics. The professor who taught from the more advanced Axler book was also a thin bearded man, and a good teacher, though I do not remember him saying anything about the number of solutions to $Ax = b$. [I do remember him saying: (i) He knew Axler; (ii) Axler's book had a slightly audacious title; (iii) Axler himself is a slightly audacious person.] I have taught several semesters of linear algebra at the University of Southern California using versions of the Strang textbook.

The *graph theory game* that I use in several exercises of the current text was an example I found online around 2008. The example was from a woman mathematician whose name and university I have unfortunately forgotten. I can no longer find the link. Whoever you are, thank you.

In the current text, the way I present the equivalent conditions for invertibility of a matrix (Theorem 5.4) is similar to what I remember from the 1990's version of Lay, Lay, and McDonald (2016). I remember that book had a perfect circle of equivalences. Here, I was only able to achieve an incomplete circle with some missing parts spliced in with if

and only if statements.

My favorite chapter of the Strang textbook is the chapter on determinants. However, one day I realized that Strang's three axioms for determinants, which are fine and good for the field $\mathbb{R}$, do not work the same way for a general field $\mathbb{F}$. That motivated me to type up the determinants chapter for a general field and with modified axioms (Chapter 11).

My style of writing is likely the closest to the Axler text. At least, I would say so. Therefore, I am surprised that these chapters I have typed up seem to have many pages, in contrast to the slim yellow book by Axler. Hopefully the added detail is helpful.

I enjoyed thinking through the steps to develop the different aspects of linear algebra. I converged on using the well known fact that a wide matrix must have a nontrivial null space to develop dimension theory (I call this the *wide matrix theorem*, Theorem 5.1). I wanted to prove Theorem 5.1 without using Gaussian elimination concepts, but I could not figure out how. So I ended up typing a standard paragraph proof that uses concepts of pivots and free variables. I didn't remember the slim yellow book discussing Gaussian elimination at all. So eventually I was curious enough to see how dimension theory was treated in that book. I found that Axler used an (of course) elegant technique of vector replacement. I kept my paragraph on free variables but, as an alternative proof, I added Exercise 5.46, which is a minor variation on the vector replacement idea.

I used to teach that inner products were linear in the first slot, until I realized the standard inner product $\langle x, y\rangle$ for $\mathbb{C}^n$ would need to be written $y^H x$ instead of $x^H y$. So I started teaching that inner products are linear in the second slot (both definitions are standard). Since real inner products are linear in *both* slots, this distinction only makes a difference for complex inner product spaces. So I wrote up Chapter 10 to provide the corresponding formulas for complex inner products (Chapter 8 treats real inner products). In general, once a student learns real inner products, the complex inner product is not much different.

I spent several days trying to prove the well known Spectral Theorem for the general case of normal matrices. The result did not seem to follow from the invariant subspace argument I was using for real and symmetric matrices. I finally figured out, while working at a Starbucks in downtown Los Angeles, that if A is normal and has eigenpair (λ, v), then A^H is normal and has eigenpair $(\overline{\lambda}, v)$ (Lemma 15.4). For me, this was the key to moving forward. I excitedly typed up the proof using this method, thinking I was giving a nice and perhaps new exposition. I then consulted the slim yellow book to find that Axler's development of the Spectral Theorem was essentially the same. While reading Axler's chapter on that topic, I learned some interesting additional properties of normal matrices. I documented those by writing the section "Additional properties of normal matrices" (Chapter 15.12). The proofs I give there are slightly different from Axler's.

This text started when I began typing up notes on projections in a *different* Starbucks in

downtown Los Angeles. I wanted to collect my thoughts on the topic, and also give students a written document of what I was teaching in class (since that material was not in the Strang book I was using). That turned into the projections chapter (Chapter 9). Somehow the mathematical justifications I needed to establish the results in the projections chapter turned into Chapters 2-8. I had already typed up notes on fields that I had been using for several years, so I added that for Chapter 1.

I was motivated to type up the chapter on eigenvalues (Chapter 12) because I wanted to see how far that theory could go with a general field $\mathbb{F}$. While doing a back-of-the-envelope calculation about an equivalent condition for affine sets, I realized the condition I was trying to prove was in fact *false* for fields that satisfy $1 + 1 = 0$. That motivated me to type up Appendix C, where I documented the counter-example (mainly for my own interest).

A student in my Fall 2026 class noticed that the axioms for a field that I presented in class (and give in Chapter 1) differ from those he had seen elsewhere: The other axioms *assumed* $0 \neq 1$ while I *proved* $0 \neq 1$. So be it. This is a cosmetic difference that does not significantly change anything. I prefer to view $0 \neq 1$ as a theorem rather than an assumption.

I would like to thank the following coffee shops in South Pasadena and downtown Los Angeles where I have done back-of-the-envelope calculations and typed up the results: Charlie's (now called Ñoños), Starbucks (multiple locations), Kafana, Coffee Bean, Philz.

I would like to thank the students and teaching assistants at the University of Southern California who have helped to spot typos and mistakes in early versions of this text.